Noname manuscript No.
(will be inserted by the editor)

Measurement Without Validity: The Compounding Reliability Problem in Agentic AI Evaluation

Preprint. Submitted to a peer-reviewed journal.

William Caban

Abstract Agentic AI systems are evaluated using automated benchmarks whose scores justify deployment decisions, safety certifications, and regulatory compliance claims. We present an empirical analysis demonstrating that these scores are systematically less trustworthy than current practice acknowledges.

The problem operates at three compounding layers. First, tasks are increasingly generated by language models: audits of ten popular benchmarks found validity flaws in seven and reporting gaps in all ten (Zhu et al, 2025). Second, human users are replaced by LLM simulators, but calibration studies document inter-simulator variance up to 9 percentage points and systematic directional miscalibration, particularly for non-Standard American English speakers (Seshadri et al, 2026). Third, our structured survey of 55 papers finds that approximately 82% apply structurally mismatched, incomplete, or absent inter-rater reliability (IRR) metrics.

These failures compound multiplicatively rather than additively. Under independence, a pipeline retaining 70% of valid signal at task generation, 80% at simulation, and 65% at judgment is at most 36% valid against the intended construct; the bound spans 0.22–0.54 across the empirical estimate range. We formalize this as $V_{\text{total}} \leq V_1 \times V_2 \times V_3$ and show it tightens further under correlated failures when the same model family operates across all three layers.

We derive eight prescriptions grounded in psychometric science: a simulation calibration floor of $\text{ICC}(A, 1) \geq 0.70$; domain-stratified reliability thresholds ($\alpha \geq 0.67$ / 0.70 / 0.80 by consequence level); structured IRR metric selection rules based on pipeline design; and IRR as a mandatory reporting field. The measurement tools exist; the field's task is to apply them.

Keywords inter-rater reliability · agentic AI evaluation · construct validity · LLM-as-a-Judge · software measurement · benchmark quality

William Caban
Red Hat, Inc., Raleigh, North Carolina, USA
Alma Mater Europaea University, Vienna, Austria
ORCID: 0009-0008-1211-7050
E-mail: wcabanba@redhat.com

1 Introduction

Deployment decisions, safety certifications, and regulatory compliance claims for agentic AI systems are all downstream of a number that appears deceptively clean: a benchmark score. That score is the product of a pipeline (task generation, world simulation, rater judgment), and each stage of that pipeline is a potential source of systematic, non-random measurement error.

The measurement problem in agentic evaluation is not simply that individual components are imperfect. It is that the imperfections compound. A pipeline that loses 30% of valid signal at task generation, 20% at simulation, and 25% at the judgment layer does not produce evaluation that is 75% valid. Under the most optimistic independence assumption, it produces evaluation that retains less than 42% of valid signal against the intended construct. Under correlated failures (which are more likely), it is worse.

This paper makes three contributions. First, we document empirically that agentic evaluation fails three distinct psychometric validity tests simultaneously, and that the empirical evidence for each failure is published, peer-reviewed, and specific. Second, we formalize the three-layer compounding validity model and show that failures multiply rather than add—a relationship the field has not yet quantified. Third, we derive eight prescriptions from the psychometric literature that practitioners and benchmark authors can apply immediately.

We are not arguing that agentic evaluation is worthless. We are arguing that it is less valid than current practices acknowledge, that the mechanisms of invalidity are known, and that the standards for reporting and interpreting evaluation results should be substantially higher.

2 Background: The Psychometric Standard

This section introduces the two psychometric constructs that underpin the analysis: construct validity, which defines what a measurement system is supposed to capture, and inter-rater reliability, which quantifies the consistency of the scoring process. Both are necessary conditions for a valid benchmark score.

2.1 Construct Validity as the Organizing Concept

The foundational criterion for any measurement system is construct validity: whether the measurement instrument actually captures the phenomenon it claims to measure (Cronbach and Meehl, 1955). Construct validity is not a single property but a family of three evaluable conditions:

– **Construct validity** proper: does the evaluation capture the concept it names? Does a benchmark described as “measuring agentic tool use” actually measure that construct, or does it measure something easier to operationalize, like API call success rate?

- **Content validity**: does the evaluation cover the full scope of the construct, or only an easily measurable subset?
- **Predictive validity**: does the evaluation score predict real-world deployment performance?

Despite its centrality to measurement science, construct validity has received limited attention in AI evaluation, even as benchmark creation involves numerous high-stakes design decisions including target user selection, construct-to-task operationalization, and scoring metric choice (Liu et al, 2024). A systematic review of 84 academic and industry papers found that technical performance was represented in 83% of evaluated works, while human-centered evaluation appeared in only 30%, and both were combined in only 15% (Meimandi et al, 2025). Benchmark success and deployment value remain systematically disconnected.

Content validity failures are measurable and consequential. A systematic analysis of 194,955 benchmark questions mapped against the EU AI Act's taxonomy of model capabilities found that current benchmarks devote 61.6% of regulatory-relevant coverage to hallucination tendency and 31.2% to performance reliability, while capabilities central to loss-of-control scenarios (including evading human oversight, self-replication, and autonomous AI development) receive zero coverage across the entire public benchmark corpus (Prandi et al, 2025).

2.2 Inter-Rater Reliability: Metrics and Their Conditions of Use

Inter-rater reliability (IRR) quantifies the degree to which raters (human annotators, LLM judges, or Agent-as-a-Judge systems) agree on the same outcome. The appropriate IRR metric is determined by the structure of the rating design, not by convention or familiarity.

Cohen's κ (Cohen, 1960) is appropriate when exactly two fixed raters score every item using the same two-rater configuration across all items. It corrects for chance agreement using per-rater marginal distributions.

Fleiss' κ (Fleiss, 1971) generalizes to three or more raters, with the critical property that rater identity may vary per item provided the number of ratings per item is constant. It uses the pooled distribution across all raters for its chance-agreement correction. This is the correct metric for fixed-size judge panels where individual judge identity may rotate.

Krippendorff's α (Krippendorff, 2004) is the most general metric: it handles any number of raters with varying identity, supports missing data natively, and accommodates nominal, ordinal, interval, and continuous measurement scales through interchangeable distance functions. For ordinal rubrics, it uses ranked distance rather than binary mismatch, correctly weighting the severity of disagreement.

Table 1 summarizes metric selection by scenario.

Table 1 IRR metric selection by agentic evaluation scenario.

Scenario	Recommended metric	Reason
3+ LLM judges, binary, complete matrix	Fleiss' κ	Nominal data, complete, multi-rater
3+ LLM judges, 1–5 rubric, complete matrix	Krippendorff's α (ordinal)	Ordinal distance metric needed
Crowdsourced annotation with missing ratings	Krippendorff's α	Handles missingness natively
2 human annotators, categorical labels	Cohen's κ	Only valid two-rater case
Continuous scores from LLM judges	Krippendorff's α (interval)	Interval distance required
Rotating judge panel membership per task	Krippendorff's α	Varying rater identity; possible missingness

2.3 Reliability Thresholds and Their Origins

The Landis–Koch scale (Landis and Koch, 1977) is the most frequently cited threshold system in AI evaluation: $\kappa > 0.21$ as "fair," $\kappa > 0.41$ as "moderate," $\kappa > 0.61$ as "substantial." This scale was developed for behavioral science contexts with human raters and has no special authority in AI evaluation. Its widespread use as an acceptability benchmark for LLM judge agreement is a category error.

For agentic evaluation, domain-calibrated thresholds grounded in construct validity theory are more appropriate:

- **Exploratory capability evaluation**: $\alpha \geq 0.67$ (Krippendorff, 2004, pp. 241–243)
- **Safety, bias, fairness, and risk scoring**: $\alpha \geq 0.70$
- **Scoring that gates deployment, compliance, or regulatory reporting**: $\alpha \geq 0.80$

3 Layer 1: Task Generation Validity

In traditional static benchmarks, tasks and ground truth are fixed. Reliability can be validated once against a stable reference. In agentic evaluation, tasks are often dynamically generated by language models, embedding systematic biases from the generator into the evaluation distribution before a single agent response is scored.

The construct validity problem at this layer is that dynamically generated tasks may not sample the intended construct uniformly. An LLM task generator that overrepresents certain task types, difficulty levels, or surface forms creates a distribution that raters can agree on perfectly, while measuring something systematically different from the intended capability.

The empirical evidence is specific. A study evaluating 10 popular agentic benchmarks found task validity flaws in 7 of them and outcome validity flaws

in 7 of them, with all 10 showing benchmark reporting gaps (Zhu et al, 2025). Documented failures include task validity failures where do-nothing agents passed 38% of airline booking tasks, and outcome validity failures where LLM judges made arithmetic errors. Every benchmark showed at least one validity failure category; none provided sufficient evidence to support unqualified capability claims.

Task generation validity must be assessed before any IRR analysis is meaningful. A rating matrix derived from systematically biased tasks cannot be rescued by metric sophistication at the judgment layer.

4 Layer 2: Simulation Calibration

Many agentic benchmarks evaluate agents in interaction with simulated users or simulated world environments rather than real humans and real environments. This design choice is operationally sensible (real human interaction at evaluation scale is expensive and slow), but it introduces a second, distinct validity problem: whether the simulated environment faithfully proxies the real deployment context.

4.1 Empirical Evidence of Simulation Miscalibration

The most direct empirical evidence comes from Seshadri et al (2026), which tested LLM user simulation against real human users on τ-Bench retail tasks across participants in the United States, India, Kenya, and Nigeria. The findings establish simulation miscalibration as a measured fact rather than a theoretical concern:

- Agent success rates varied up to 9 percentage points across different LLM user simulators, demonstrating that simulation results are not robust to simulator choice.
- Evaluations using simulated users exhibited systematic directional miscalibration: underestimating agent performance on challenging tasks while overestimating it on moderately difficult ones.
- Simulated users introduced conversational artifacts absent from real interactions: elevated question-asking, heightened politeness markers, and artificial turn structures.

4.2 The Demographic Asymmetry Problem

The calibration failures documented by Seshadri et al (2026) are not uniformly distributed across user populations. African American Vernacular English (AAVE) speakers experienced consistently worse success rates and calibration errors than Standard American English (SAE) speakers, with disparities compounding with age. Indian English speakers showed similar patterns.

The same asymmetry extends to LLM judges evaluating across linguistic varieties. In a study evaluating three LLMs as toxicity judges across 60 language varieties spanning 10 language clusters, LLM–human agreement was the weakest dimension of consistency, weaker than cross-model or cross-dialect consistency, with systematic gaps across dialectal groups (Faisal et al, 2025). Judge miscalibration for non-standard dialect speakers persists in direct LLM judgment, independently of simulation layer failures.

This is simultaneously a measurement validity failure and a fairness failure. An evaluation pipeline calibrated on SAE simulated users is not a valid measurement of agentic capability for AAVE-speaking or non-SAE users.

4.3 Why Agreement on a Distorted Signal Is Not Reliability

The key inferential step at this layer is often missed: high rater agreement on simulated interactions does not demonstrate evaluation reliability. It demonstrates that all raters agree on a distorted input.

A scoring panel that perfectly agrees on agent scores produced by an AAVE-miscalibrated simulator has achieved high IRR in the technical sense. It has also produced a high-reliability measurement of the wrong thing. Agreement on a distorted signal is agreement on distortion.

5 Layer 3: Metric Misspecification

The third layer of validity failure occurs at judgment: even when a human or LLM judge scores agent outputs, the inter-rater reliability metric used to validate that scoring is structurally mismatched to the pipeline design in the large majority of published evaluations.

5.1 Why Cohen's κ Is Structurally Wrong for Most Agentic Evaluation Designs

Cohen's κ requires exactly two raters scoring every item, with the same two-rater pair maintained across all items. Almost no agentic evaluation design satisfies this constraint:

- LLM-as-a-Judge pipelines typically use a rotating pool of judge models that vary across evaluation runs or item subsets.
- Agent-as-a-Judge frameworks (Zhuge et al, 2025) explicitly assign different agent-judges per task type, domain, or stakeholder persona.
- Human annotation at scale uses crowdsourcing pools where rater identity varies per item.
- Dynamic benchmark generation produces item batches scored by different annotator subsets.

Applying Cohen's κ to any of these designs produces a mathematically invalid result, because its chance-agreement correction assumes fixed rater identity. Using it anyway (common in practice) is a silent validity failure that produces an uncorrectable reliability estimate without warning.

5.2 The Literature Scan: IRR in Published Agentic Evaluation Papers

Methodology. We conducted a structured survey of 55 papers published or posted between 2022 and 2026, selected via a stratified purposive approach across nine topic categories: major agentic benchmarks; automated grader validity studies; benchmarks with correct IRR use; LLM-as-a-Judge studies; benchmarks with structural metric mismatch; long-horizon evaluation; safety and RLHF annotation; recent 2025–2026 evaluation papers; and safety-critical IRR failure cases. Within each category, papers were selected for citation prominence, venue tier (ACL, EMNLP, NeurIPS, ICLR), and direct relevance to evaluation methodology; arXiv preprints in cs.AI and cs.CL supplemented peer-reviewed coverage where thin. This is a purposive stratified survey, not a systematic review: the sample is designed to cover the main failure modes rather than to exhaust a keyword-defined corpus. Coding was applied against four pre-specified dimensions: (1) the IRR metric reported or absent, (2) the rater design (number of raters, fixed or rotating identity), (3) the measurement scale (binary, nominal, ordinal, or continuous), and (4) whether the metric's structural assumptions were satisfied by the pipeline design. Coding was performed by a single rater; this limitation is addressed in Section 8. The full coding table is in Appendix A.

Finding (structured estimate; single rater). Across 55 coded papers, approximately 10 (18%) appear to use structurally correct IRR metrics with explicit rationale. The remaining 45 exhibit one of three failure modes:

Mode 1: Automated Grading, No IRR (11%). Major benchmark papers—τ-bench (Yao et al, 2024), OSWorld (Xie et al, 2024), SWE-bench (Jiménez et al, 2024), GAIA (Mialon et al, 2024), and WebArena (Zhou et al, 2024)—use deterministic automated graders or unvalidated LLM judges with no IRR measurement against human annotation; τ^2-Bench (Barres et al, 2025), the quality-fix follow-up to τ-bench, takes the same approach: tighter automated grading rather than human IRR validation. Gurram (2026) directly measures the gap: substring-based automated evaluation achieves $\kappa = 0.049$ against human annotation (essentially chance-level) while a three-LLM ensemble achieves only $\kappa = 0.432$. The Verified re-releases of these benchmarks (WebArena Verified, El Hattami et al 2025: Cohen's $\kappa = 0.83$) demonstrate that rigorous human annotation produces fundamentally different reliability evidence.

Mode 2: Structural Metric Mismatch (16%). Papers that do report IRR most commonly apply Cohen's κ to panels of three or more LLM judges, or report raw percentage agreement for ordinal and ternary scales without chance correction

(Fan et al 2026 AgentProcessBench: 89.1% on ternary labels). In Han et al (2025) the mismatch is more subtle: Cohen's κ is applied as a series of pairwise comparisons (each of 54 LLMs vs. human ratings), each individually a valid two-rater design, but the results are then aggregated and interpreted as panel reliability across 54 configurations with varying rater compositions, a use case requiring Fleiss' κ or Krippendorff's α, not repeated two-rater κ.

Mode 3: Wrong Metric for the Measurement Construct (31%). Safety, preference, and RLHF papers apply metrics appropriate for one statistical question to answer a different one: Pearson correlation for ordinal safety ratings from 112 demographically diverse annotators (Movva et al, 2024); ELO ratings substituting for inter-annotator reliability in crowdsourced pairwise preference (Chiang et al, 2024); raw percent agreement without chance correction for 40 contractors ranking model outputs (Ouyang et al, 2022, 72–77%).

Table 2 summarizes the distribution.

Table 2 IRR failure modes across 55 coded papers. Only 10 (18%) use structurally correct metrics with explicit rationale.

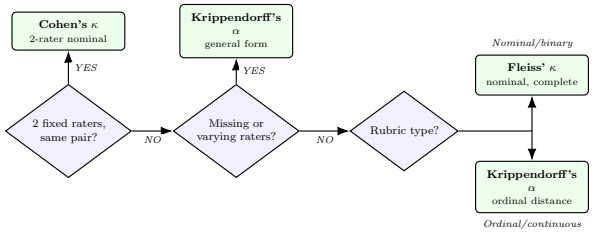
Failure mode	Papers	%	Representative example
Automated grading, no IRR	6	11	τ-bench, OSWorld, SWE-bench
Structural metric mismatch	9	16	Cohen's κ for 54-LLM panel (Han et al, 2025)
Wrong construct / % only	17	31	InstructGPT; Chatbot Arena (ELO)
No metric reported	13	24	RLHF pipelines, Red Teaming
Correct metric	**10**	**18**	WebArena Verified ($\kappa = 0.83$)

Structural contrast. The 10 papers using IRR correctly share a distinguishing behavior: they report multiple metrics explicitly and state their rationale. Papers using incorrect metrics report only Cohen's κ with no structural justification. The mismatch is not a deliberate simplification; it reflects unawareness of the structural conditions that govern metric validity.

5.3 Ordinal Rubrics and the Kappa Penalty

Most agentic evaluation rubrics use ordinal scoring (e.g., 1–5 on helpfulness, safety, coherence). Fleiss' κ treats all disagreements as categorically equal: a rater scoring 1 when others score 5 is penalized identically to a rater scoring 4 when others score 5. This is wrong for ordinal data.

The consequence is bidirectional distortion: Cohen's κ and Fleiss' κ artificially inflate apparent disagreement when raters are closely aligned on an ordinal scale (one point apart) and can underestimate disagreement when raters are at opposite ends of the scale. For safety evaluation, this matters: a safety rubric where judges disagree by 1 point appears as unreliable as one where they disagree by 4 points.



Selection rules:

Cohen's κ
Exactly 2 fixed raters, same pair, every item.

Fleiss' κ
$\geq$3 raters, complete matrix, nominal/binary rubric.

Krippendorff's α (general)
Missing ratings or rater identity varies.

Krippendorff's α (ordinal)
Complete matrix; ordinal or continuous rubric.

Using Cohen's κ outside the 2-rater fixed-pair case produces an invalid reliability estimate without warning.

Fig. 1 IRR metric selection. Left: decision tree based on rater design, matrix completeness, and rubric scale. Right: selection rules summary. Using Cohen's κ outside the two-rater fixed-pair case is a silent validity failure.

5.4 Metric Selection Decision Tree

Figure 1 provides a decision tree for selecting the structurally correct IRR metric based on rater design, data completeness, and measurement scale.

6 The Compounding Argument

The full mathematical derivation, including the correlated-failure extension and a practical estimation protocol, is in Appendix B.

6.1 Formal Statement

Let $V_1 \in [0, 1]$ denote the task generation validity: the degree to which the generated task distribution matches the ideal construct distribution, defined as one minus the total variation distance between them. Because the ideal construct distribution $D_{\mathcal{C}^*}$ is a theoretical object and not directly observable, V_1 cannot be computed from first principles; in practice it is estimated by comparing generated tasks against a curated expert reference set (Appendix B, §B.6). Let $V_2 \in [0, 1]$ denote the simulation calibration validity, defined as the ICC$(A, 1)$ between simulated and real agent outcomes on a matched task set, estimated separately per user population group. Let $V_3 \in [0, 1]$ denote the judgment validity: the appropriate IRR metric value (Krippendorff's α or Fleiss' κ, per the pipeline's measurement scale) for the rating design.

The overall construct validity of the evaluation pipeline is bounded above by:

$$V_{\text{total}} \leq V_1 \times V_2 \times V_3 \tag{1}$$

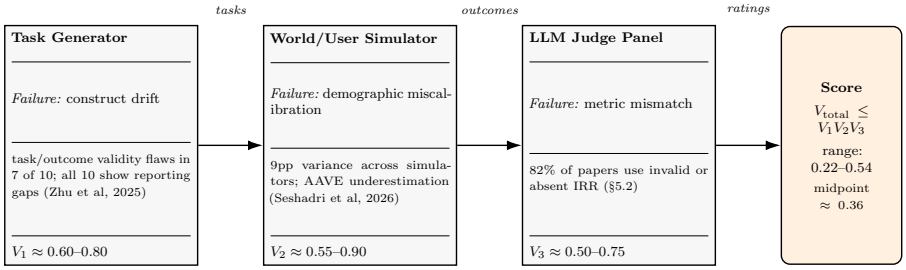


Fig. 2 Three-layer agentic evaluation pipeline with validity estimates. Each layer introduces a distinct failure; failures compound multiplicatively under the independence bound ($V_{\text{total}} \leq V_1 \times V_2 \times V_3$, range 0.22–0.54, midpoint ≈ 0.36). The bound tightens further under correlated failures when the same model family operates at all three layers.

The bound holds under the assumption of independence between layer failures. When layer failures are positively correlated (as they are likely to be when the same provider family operates at all three layers), V_{total} falls below this bound. The independence assumption is therefore optimistic; the correlated-failure bound is tighter (Appendix B, §B.4).

Figure 2 illustrates the three-layer pipeline with its validity estimates and the compounding formula.

6.2 Worked Example

Using empirical estimates from the cited literature:

- V_1 **(task generation)**: Zhu et al (2025) found task validity flaws in 7 of 10 popular benchmarks and outcome validity flaws in 7 of 10, with all 10 showing benchmark reporting gaps. For a typical dynamic benchmark, $V_1 \approx 0.60$–0.80.
- V_2 **(simulation)**: Seshadri et al (2026) documented 9pp variance in success rates across simulators and systematic directional miscalibration against real human users. The V_2 ranges below are expert-judgment estimates anchored to the documented magnitude of directional miscalibration, not computed values. Conservative estimate: $V_2 \approx 0.75$–0.90 for SAE populations; $V_2 \approx 0.55$–0.75 for AAVE or Indian English populations.
- V_3 **(judgment)**: Zhou et al (2026) found significant and diverse bias patterns in LLM judges. With within-family judge panels and ordinal rubrics scored with Fleiss' κ, $V_3 \approx 0.50$–0.75.

Taking the empirical midpoint of each range ($V_1 \approx 0.70$; $V_2 \approx 0.80$, the SAE estimate; $V_3 \approx 0.65$) gives a plausible midpoint scenario:

$$V_{\text{total}} \leq 0.70 \times 0.80 \times 0.65 \approx 0.36 \tag{2}$$

Across the full range ($V_1 \in [0.60, 0.80]$, $V_2 \in [0.75, 0.90]$ for SAE, $V_3 \in [0.50, 0.75]$), the bound spans $V_{\text{total}} \in [0.225, 0.54]$. For a non-SAE user popula-

tion, substituting the non-SAE V_2 midpoint ($V_2 \approx 0.65$) gives:

$$V_{\text{total}} \leq 0.70 \times 0.65 \times 0.65 \approx 0.30 \tag{3}$$

An evaluation pipeline that appears to be functioning can be retaining less than 30% of valid signal for the user populations that LLM simulators systematically underserve.

6.3 Implications for Published Benchmark Claims

When a paper reports that "Agent X achieves 82% on [benchmark]," the epistemic weight of that claim is conditioned on the validity of the three pipeline layers. Under the compounding model, an 82% score from an evaluation with $V_{\text{total}} \approx 0.36$ does not tell us that the agent achieves 82% of the intended capability. It tells us that the agent achieves some unknown capability level, bounded by a measurement apparatus that is less than 40% valid.

This does not mean the score is uninformative. It means its interpretation requires the three validity layers to be assessed and reported alongside the score, as a condition of the score having evidentiary weight.

7 Prescriptions for Valid Agentic Evaluation

Items supported by existing literature are stated as requirements; items that extend beyond current literature are marked *we propose*.

1. **Validate simulation calibration before trusting evaluation results.** Use a held-out set of real human interactions to measure the ICC$(A, 1)$ between simulated and real agent outcomes. If calibration is not validated, the evaluation signal is of unknown reliability regardless of how precisely the judging rubric is specified (Seshadri et al, 2026). Report the calibration ICC alongside any benchmark result that uses user simulation. *We propose ICC$(A, 1) \geq 0.70$ as a minimum threshold for simulation calibration, set 3pp above Krippendorff (2004)'s general content analysis minimum of 0.67 because LLM simulators exhibit systematic behavioral artifacts absent from trained-human rater contexts (Seshadri et al, 2026).*
2. **Select the IRR metric based on pipeline structure, not convention.** Apply Cohen's κ only when exactly two fixed raters score every item. Apply Fleiss' κ when a fixed-size panel of three or more raters scores nominal or categorical outcomes on a complete matrix. Apply Krippendorff's α when the rubric is ordinal or continuous, when missing ratings are possible, or when rater panel membership varies. Using Cohen's κ outside its structural assumptions is a silent validity failure (Section 5).
3. **Report both Fleiss' κ and Krippendorff's α for any rubric using ordinal or interval scales.** When the two metrics diverge, the divergence is itself diagnostic: $\kappa < \alpha$ suggests raters are frequently one scale point apart

rather than at opposite ends of the range, a much healthier disagreement pattern than the κ score alone would imply.

4. **Use cross-family judge ensembles.** Self-preference and family-level bias in LLM judges (Spiliopoulou et al, 2025; Wataoka et al, 2024) are large enough to distort comparative agent rankings. Evaluation pipelines should use judges from at least two distinct provider families and report inter-judge agreement across families. Single-family judge panels produce spuriously inflated within-family agreement.
5. **Apply domain-appropriate reliability thresholds.** For exploratory capability evaluation, $\alpha \geq 0.67$ is the minimum consistent with Krippendorff (2004, pp. 241–243)'s own recommendation. For safety, bias, fairness, and risk scoring, *we propose* $\alpha \geq 0.70$. For scoring that gates model deployment, compliance certification, or regulatory reporting, *we propose* $\alpha \geq 0.80$. When these thresholds cannot be met, as Jafari et al (2026) demonstrate for LLM mental health response evaluation (three certified psychiatrists produced ICC = 0.087–0.295 and $\alpha = -0.203$), the correct response is not to lower the threshold but to recognize that the construct is insufficiently specified for reliable measurement.
6. **Validate dynamically generated task distributions against curated reference sets.** When tasks are generated by an LLM, validate that the generated distribution samples the intended construct by testing against a curated reference set scored by domain experts (Zhu et al, 2025). Report the total variation distance or KL divergence between generated and reference distributions.
7. **Stratify evaluation by demographic and linguistic group.** Simulation calibration failures are not uniformly distributed across user populations (Seshadri et al, 2026), and LLM judge calibration failures show the same pattern across dialect groups (Faisal et al, 2025). AAVE speakers, Indian English speakers, and other non-SAE populations must be included in calibration validation before benchmark results are considered representative.
8. **Report IRR as a required field in evaluation documentation.** Standardized reporting frameworks such as EvalCards (Dhar et al, 2025) provide the infrastructure to enforce consistent disclosure across papers and model releases. Including IRR metric, threshold, and metric-selection rationale as required fields in evaluation reporting cards would institutionalize the prescriptions above at the submission level.

8 Threats to Validity

We assess threats to the validity of this study following the threat classification standard in empirical software engineering research (Wohlin et al, 2012; Jedlitschka et al, 2008).

8.1 Construct Validity

Operationalization of V_1, V_2, V_3. The three validity measures are operationalizations of theoretical constructs. The ideal construct distribution $D_{\mathcal{C}^*}$ (Section 6.1) is a theoretical object; V_1 cannot be computed from first principles and must be estimated via expert-curated reference sets. V_2 is defined as $\text{ICC}(A, 1)$ between simulated and real outcomes, but the cited evidence (Seshadri et al 2026) reports inter-simulator variance, not direct $\text{ICC}(A, 1)$ values; the V_2 ranges used in Section 6.2 are expert-judgment estimates, not computed quantities. Readers should treat the numerical ranges as illustrative rather than precise.

IRR coding taxonomy. The four-dimension coding scheme (Section 5.2) may not capture all relevant failure modes. In particular, papers that apply the structurally correct metric but with an incorrect threshold, or that report IRR on a subset of items not representative of the full evaluation, fall outside the coding scheme and may be classified as "correct" when they should not be. Appendix A provides the full coding table to support independent assessment.

Compounding model. The multiplicative validity bound (Equation 1) is presented as a conceptual model, explicitly not a proved theorem (Appendix B, §B.3). The motivating argument assumes layer independence and treats each V_i as a normalized signal-to-noise ratio; both assumptions are approximations. The bound should be read as a formalization of the compounding intuition, not a mathematical guarantee.

8.2 Internal Validity

Single-rater coding. The literature scan (Section 5.2) was coded by a single rater without a second rater for IRR measurement. This creates a self-referential tension: the paper's central empirical claim (82% IRR misuse) rests on exactly the methodological choice the paper argues against. By the paper's own standard ($\alpha \geq 0.67$ for exploratory evaluation), the 82% statistic cannot meet the minimum reliability threshold because the coding design has one rater and α is undefined for a single rater. The 82% figure is therefore a structured expert estimate, not an IRR-validated prevalence. The pre-specified coding criteria and full coding table in Appendix A make individual coding decisions verifiable in lieu of replication.

Purposive sampling and confirmation bias. The 55-paper sample was selected purposively to cover nine topic categories, not to exhaust a keyword-defined corpus. This exposes the scan to selection bias: papers demonstrating failures may be more visible (cited more, identified more easily) than papers with correct but unreported IRR practices. The author's prior belief that IRR is systematically misused may have influenced which papers were included in each category or how ambiguous cases were coded. The structured category definitions and transparent category membership (Appendix A) mitigate but do not eliminate this threat.

8.3 External Validity

Sample scope. The 55 papers were drawn primarily from English-language venues (ACL, EMNLP, NeurIPS, ICLR) and arXiv preprints in cs.AI and cs.CL. Papers from other geographic regions, language communities, or non-English venues may show different IRR reporting patterns. The temporal scope (2022–2026) means that earlier work and work after the scan cutoff is not captured; IRR practices are evolving and the field may improve.

Domain of the compounding model. The three-layer model was constructed for agentic evaluation pipelines involving task generation, user simulation, and LLM judgment. Static benchmarks with fixed tasks, no simulation layer, and human-only raters present a different threat model where only Layer 3 (IRR metric selection) directly applies. The prescriptions in Section 7 are intended to generalize across evaluation pipeline designs, but their applicability to static benchmarks should be assessed separately.

8.4 Conclusion Validity

Independence assumption. The primary conclusion of Section 6 is that validity failures compound multiplicatively. This conclusion depends on the independence assumption: that layer failures are statistically unrelated. When the same underlying mechanism (e.g., the same LLM provider family) operates at all three layers, failures are positively correlated and V_{total} falls below the independence bound. The conclusion is therefore stated conservatively: the independence bound is presented as optimistic, and correlated-failure scenarios are shown to produce lower V_{total} (Appendix B, §B.4). The qualitative conclusion (failures multiply rather than add) is robust to the independence assumption because the bound holds as an upper limit regardless of correlation structure.

Empirical range estimates. The V_i ranges used in the worked example (Section 6.2) are derived from the cited literature via expert judgment. The conclusion that "less than 40% of valid signal" is possible under a midpoint scenario is sensitive to these ranges. A reader who assigns different ranges will obtain a different midpoint. The sensitivity of V_{total} to the input ranges is implicit in the 0.22–0.54 span reported, which reflects the full cross-product of the documented empirical estimate ranges.

9 Discussion

9.1 What This Paper Does Not Claim

We do not claim that current agentic evaluations are worthless. They provide directional signal. We claim that the field's current practices overstate the validity and precision of that signal, and that the standards for reporting should

be substantially higher. A benchmark result reported without IRR metrics, simulation calibration evidence, and metric selection justification is incomplete, in the same way that a clinical measurement reported without instrument validity data is incomplete.

9.2 The Governance Stakes

For academic benchmark comparisons, validity limitations are a scientific quality issue. For deployment decisions, safety certifications, and regulatory claims, they are a governance failure. Agentic AI systems are increasingly being deployed in consequential contexts. The evaluation apparatus used to justify those deployments must meet the same validity standards applied to any measurement-driven governance process.

The regulatory dimension makes this concrete. Emerging frameworks such as the EU AI Act require evidence of comprehensive risk assessment for general-purpose AI models. A systematic analysis of 194,955 benchmark questions found that capabilities central to loss-of-control (including evading human oversight, self-replication, and autonomous AI development) receive zero benchmark coverage (Prandi et al, 2025). If the construct coverage gap means that no existing benchmark can provide evidence of regulatory compliance, then the IRR validity of those benchmarks is a secondary concern: the primary problem is that they do not measure what regulations require. Both failures must be addressed together.

9.3 Broader Impact

This work raises the evidential bar for deployment certifications, surfaces demographic calibration failures that existing evaluation practice obscures, and gives practitioners eight actionable criteria they can apply using existing psychometric tools.

Three misuse risks follow from this work. First, the proposed reliability thresholds ($\alpha \geq 0.70$, $\alpha \geq 0.80$) may become compliance checkboxes rather than genuine validity indicators. Meeting a threshold does not guarantee valid measurement; it establishes the minimum condition for a score to be interpretable. Second, stratified calibration across demographic and linguistic groups increases evaluation cost; benchmark maintainers and shared evaluation platforms should bear primary responsibility for calibration coverage, not individual research teams. Third, the finding that three certified psychiatrists produced ICC = 0.087–0.295 on LLM mental health safety evaluation (Jafari et al, 2026) should not discourage research in high-stakes AI domains. It means the construct is underspecified for reliable measurement, which calls for more rigorous construct development, not withdrawal.

10 Conclusion

The move from static to dynamic agentic evaluation does not reduce the importance of inter-rater reliability; it multiplies the surfaces on which reliability can fail and introduces new systematic biases at each layer: task generation, world simulation, and judgment. The compounding effect means that an evaluation pipeline with moderate validity problems at each of three layers can produce an evaluation signal that is far less valid than any single-layer estimate would suggest.

The science of measurement was not built for static benchmarks alone. It was built for situations where the measurement apparatus is a source of variance, where raters disagree in structured rather than random ways, and where the construct must be separated from the instrument measuring it. That is agentic evaluation.

These standards are achievable. El Hattami et al (2025) (WebArena Verified) demonstrates this directly: adding rigorous human annotation with two fixed annotators and reporting Cohen's $\kappa = 0.83$ [0.81, 0.85] satisfies the metric selection requirement (Prescription 2), exceeds the exploratory reliability threshold (Prescription 5), and treats IRR as a required reporting field (Prescription 8). Three of the eight prescriptions, met by a design choice the field already knows how to make.

Fleiss' κ, Krippendorff's α, ICC, and construct validity frameworks are available, validated, and applicable. The field's task is to use them, to report IRR as a first-class metric alongside every benchmark result, and to recognize that a score without validity evidence is not a measurement. It is a number.

Acknowledgements The author thanks the agentic AI evaluation community whose published work—including the benchmark papers, LLM-as-a-Judge studies, and calibration analyses cited throughout—made this synthesis possible.

Statements and Declarations

Competing Interests. The author is employed by Red Hat, Inc. This research was conducted independently as part of the author's doctoral program at Alma Mater Europaea University. The views expressed are the author's own and do not represent the official position of Red Hat, Inc. The author declares no financial competing interests.

Author Contributions. Single-author paper. William Caban conceived the study, conducted the literature scan, developed the compounding validity model, and wrote the manuscript.

Use of AI Writing Assistance. The author used LLM assistance (Claude, Anthropic) to assist with manuscript preparation, specifically: converting the LaTeX template format to SVJour3 and editing text for length. All intellectual

content, analytical claims, literature survey findings, and conclusions are the author's own. The author reviewed and is responsible for all AI-assisted text. Claude is not listed as an author.

Funding. No funding was received for this research.

Ethics Approval. This study is based entirely on analysis of publicly available published research. No human participants, personal data, or animal subjects were involved.

Data Availability

The literature scan coding table (Appendix A) constitutes the primary data generated by this study. All 55 papers included in the scan are publicly available at the venues and arXiv identifiers cited. The coding criteria, category definitions, and classification scheme are fully documented in Appendix A, enabling independent replication.

No proprietary datasets, model weights, or benchmark systems were created or used. Code for the compounding model (Appendix B) requires no implementation beyond the equations stated.

References

Barres V, Dong H, Ray S, Si X, Narasimhan K (2025) τ^2-bench: Evaluating conversational agents in a dual-control environment. arXiv preprint arXiv:2506.07982

Chiang WL, Zheng L, Sheng Y, Angelopoulos AN, Li T, Li D, Zhang H, Zhu B, Jordan M, Gonzalez JE, Stoica I (2024) Chatbot arena: An open platform for evaluating LLMs by human preference. arXiv preprint arXiv:2403.04132

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37–46, DOI 10.1177/001316446002000104

Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychological Bulletin 52(4):281–302, DOI 10.1037/h0041570

Dhar R, Villegas DS, Karamolegkou A, Schiavone A, Yuan Y, Chen X, Li J, Frank S, De Grazia L, Swain M, et al (2025) EvalCards: A framework for standardized evaluation reporting. arXiv preprint arXiv:2511.21695

El Hattami A, Thakkar M, Chapados N, Pal C (2025) WebArena Verified: Reliable evaluation for web agents. In: Scaling Environments for Agents Workshop, NeurIPS 2025, URL `https://openreview.net/forum?id=94tlGxmqkN`

Faisal F, Rahman MM, Anastasopoulos A (2025) Dialectal toxicity detection: Evaluating LLM-as-a-judge consistency across language varieties. In: Findings of the Association for Computational Linguistics: EMNLP 2025, arXiv:2411.10954

Fan S, Ye X, Huo Y, Chen ZY, Guo Y, Yang S, Yang W, Ye S, Chen J, Chen H, Cong X, Lin Y (2026) AgentProcessBench: Diagnosing step-level process quality in tool-using agents. arXiv preprint arXiv:2603.14465, under review

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5):378–382, DOI 10.1037/h0031619

Gurram B (2026) Evaluating tool-using language agents: Judge reliability, propagation cascades, and runtime mitigation in AgentProp-Bench. arXiv preprint arXiv:2604.16706, under review

Han S, Titericz Junior G, Balough T, Zhou W (2025) Judge's verdict: A comprehensive analysis of LLM judge capability through human agreement. arXiv preprint arXiv:2510.09738

Jafari K, Rust PUN, Eddy D, Fraser R, Vasan N, Djordjevic D, Dadlani A, Lamparth M, Kim E, Kochenderfer MJ (2026) Expert evaluation and the limits of human feedback in mental health AI safety testing. arXiv preprint arXiv:2601.18061, under review

Jedlitschka A, Ciolkowski M, Pfahl D (2008) Reporting experiments in software engineering. In: Shull F, Singer J, Sjøberg DI (eds) Guide to Advanced Empirical Software Engineering, Springer, pp 201–228, DOI 10.1007/978-1-84800-044-5_8

Jiménez CE, et al (2024) SWE-bench: Can language models resolve real-world GitHub issues? arXiv preprint arXiv:2310.06770

Krippendorff K (2004) Content Analysis: An Introduction to Its Methodology, 2nd edn. Sage Publications, Thousand Oaks, CA

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174, DOI 10.2307/2529310

Liu YL, Blodgett SL, Cheung JCK, Liao QV, Olteanu A, Xiao Z (2024) ECBD: Evidence-centered benchmark design for NLP. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 16349–16365, DOI 10.18653/v1/2024.acl-long.861

Meimandi KJ, Aránguiz-Dias G, Kim GR, Saadeddin L, Griffith A, Kochenderfer MJ (2025) The measurement imbalance in agentic AI evaluation undermines industry productivity claims. arXiv preprint arXiv:2506.02064

Mialon G, et al (2024) GAIA: A benchmark for general AI assistants. arXiv preprint arXiv:2311.12983

Movva R, Koh PW, Pierson E (2024) Annotation alignment: Comparing LLM and human annotations of conversational safety. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, arXiv:2406.06369

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, et al (2022) Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems, vol 35, pp 27730–27744, arXiv:2203.02155

Prandi M, Suriani V, Pierucci F, Galisai M, Nardi D, Bisconti P (2025) Bench-2-CoP: Can we trust benchmarking for EU AI compliance? arXiv preprint arXiv:2508.05464

Seshadri P, Cahyawijaya S, Odumakinde A, Singh S, Goldfarb-Tarrant S (2026) Lost in simulation: LLM-simulated users are unreliable proxies for human users in agentic evaluations. arXiv preprint arXiv:2601.17087, under review

Spiliopoulou E, Fogliato R, Burnsky H, Soliman T, Ma J, Horwood G, Ballesteros M (2025) Play favorites: A statistical method to measure self-bias in LLM-as-a-judge. arXiv preprint arXiv:2508.06709

Wataoka K, Takahashi T, Ri R (2024) Self-preference bias in LLM-as-a-judge. In: NeurIPS 2024 Safe Generative AI Workshop, arXiv:2410.21819

Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2012) Experimentation in Software Engineering. Springer, DOI 10.1007/978-3-642-29044-2

Xie T, et al (2024) OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. arXiv preprint arXiv:2404.07972

Yao S, et al (2024) τ-bench: A benchmark for tool-agent-user interaction in real-world domains. arXiv preprint arXiv:2406.12045

Zhou H, Huang H, Zhang R, Chen K, Xu B, Zhu C, Zhao T, Yang M (2026) Toward robust LLM-based judges: Taxonomic bias evaluation and debiasing optimization. arXiv preprint arXiv:2603.08091, under review

Zhou S, Xu FF, Zhu H, Zhou X, Lo R, Sridhar A, Cheng X, Bisk Y, Fried D, Alon U, Neubig G (2024) WebArena: A realistic web environment for building autonomous agents. In: The Twelfth International Conference on Learning Representations, URL `https://openreview.net/forum?id=oKn9c6ytLx`

Zhu Y, Jin T, Pruksachatkun Y, Zhang A, Liu S, Cui S, Kapoor S, Longpre S, Meng K, Weiss R, et al (2025) Establishing best practices for building rigorous agentic benchmarks. arXiv preprint arXiv:2507.02825

Zhuge M, Zhao C, Ashley DR, Wang W, Khizbullin D, Xiong Y, Liu Z, Chang E, Krishnamoorthi R, Tian Y, Shi Y, Chandra V, Schmidhuber J (2025) Agent-as-a-judge: Evaluate agents with agents. In: Proceedings of the 42nd International Conference on Machine Learning, arXiv:2410.10934

A Literature Scan: Coding Criteria and Category Summary

This appendix documents the methodology and results of the structured literature scan reported in Section 5.2. It presents the pre-specified coding criteria, a category summary of the 55 coded papers, and notable individual cases referenced in the main text.

A.1 Pre-Specified Coding Criteria

Each of the 55 papers was coded on four dimensions, applied uniformly before reading the paper's results section:

1. **IRR metric reported**: Cohen's κ, Fleiss' κ, Krippendorff's α, ICC (form specified or unspecified), percentage agreement only, or no metric reported.
2. **Rater design**: number of raters; whether rater identity was fixed across items or varied (rotating pool, crowdsourced, or multiple independent runs of the same model).
3. **Measurement scale**: binary, nominal (unordered categories), ordinal or ternary (ordered but discrete), or continuous.

4. **Structural validity**: whether the reported metric's structural assumptions were satisfied by the rater design and measurement scale. Coding applied the following rules: Cohen's κ for > 2 raters or rotating identity $\rightarrow$ MISMATCH; Fleiss' κ for ordinal or continuous scale $\rightarrow$ PARTIAL MISMATCH; percentage agreement without chance correction $\rightarrow$ INCOMPLETE; no IRR with automated grader of unvalidated accuracy $\rightarrow$ LAYER-3 VALIDITY FAILURE.

A.2 Category Summary

Table 3 summarizes the 55 papers by topic category.

Table 3 Literature scan category summary. "Correct" = structurally valid metric with explicit rationale. "Failure mode" = dominant pattern among incorrect papers.

Cat.	Topic	Papers	Correct	Dominant failure mode
A	Major agentic benchmarks (automated grading)	6	0	No IRR; automated grader validity assumed
B	Automated grader validity studies	4	4	— (all correct; measuring the gap)
C	Benchmarks with formal IRR	4	4	— (all correct)
D	LLM-as-a-Judge studies	6	0	Cohen's κ for multi-model panels
E	Benchmarks with structural metric mismatch	7	0	Cohen's κ for ordinal rubrics; % only
F	Long-horizon and capability benchmarks	6	0	No metric or % agreement only
G	Safety, RLHF, and preference evaluation	6	0	Pearson/ELO for ordinal IRR problems
H	Recent 2025–2026 evaluation papers	5	2	Metric stated; structural rationale absent
I	Safety-critical IRR failures	11	0*	Raw % agreement; no metric; wrong metric
Total		**55**	**10 (18%)**	

* Category I includes one paper (Jafari et al 2026) that uses the correct metrics but reports catastrophically low values (ICC = 0.087–0.295, $\alpha = -0.203$).

A.3 Notable Individual Cases

- Jafari et al (2026): Three certified psychiatrists, correct metrics, ICC = 0.087–0.295 and $\alpha = -0.203$ on LLM mental health response safety: highest-stakes domain, most disagreement.
- Ouyang et al (2022): 40 rotating contractors, 4-way preference ranking. Reports 72–77% raw percent agreement with no chance correction; requires Krippendorff's α for rotating-rater ordinal design.
- Gurram (2026): Direct grader-vs-human comparison; substring grading achieves $\kappa = 0.049$ (chance-level), 3-LLM ensemble $\kappa = 0.432$.
- El Hattami et al (2025) (WEBARENA VERIFIED): Two annotators per task (primary + verifier), binary success/fail labels, Cohen's $\kappa = 0.83$ [0.81, 0.85] across all 812 tasks: correct metric, threshold exceeded, reporting complete. The strongest positive example in the scan.

B Formal Derivation of the Compounding Validity Model

This appendix provides the mathematical treatment supporting Section 6.

B.1 Definitions

Let an agentic evaluation pipeline Π consist of three sequential components:
- G: a task generation function that samples tasks t from a distribution D_G over task space $\mathcal{T}$
- S: a simulation function that, given a task t and an agent A, generates an interaction trajectory $\tau = S(t, A)$ drawn from a distribution $D_S(t)$
- J: a judgment function that, given a trajectory τ, assigns a score $s = J(\tau) \in \mathbb{R}$

Let $\mathcal{C}^*$ denote the target construct: the latent capability the evaluation is designed to measure. The construct $\mathcal{C}^*$ induces an ideal task distribution $D_{\mathcal{C}^*}$ over $\mathcal{T}$ and an ideal scoring function J^* that maps agent behavior to true capability.

B.2 Layer-Specific Validity Measures

V_1: Task Generation Validity.

$$V_1 = 1 - d_{\mathrm{TV}}(D_G,\, D_{\mathcal{C}^*})$$

where d_{TV} denotes total variation distance. $V_1 = 1$ when $D_G = D_{\mathcal{C}^*}$ exactly; $V_1 = 0$ when the supports are disjoint.

V_2: Simulation Calibration Validity.

$$V_2 = \mathrm{ICC}(\mathrm{outcomes}_{\mathrm{sim}},\, \mathrm{outcomes}_{\mathrm{real}})$$

where ICC is computed using the two-way mixed-effects model for absolute agreement ($\mathrm{ICC}(A, 1)$). V_2 must be estimated separately per demographic and linguistic user group:

$$V_2 = \mathbb{E}_{g \sim P_{\mathrm{users}}}[V_2^{(g)}]$$

V_3: Judgment Validity. V_3 measures the degree to which the rating protocol produces valid inter-rater agreement, using the appropriate IRR metric for the pipeline's measurement scale.

B.3 The Multiplicative Validity Bound

Conceptual bound. We state the following as a conceptual model, not a proved mathematical proposition. Under the assumption that V_1, V_2, V_3 are independently determined:

$$V_{\mathrm{total}}(\Pi) \leq V_1 \cdot V_2 \cdot V_3$$

Motivating argument. Let q denote an agent query drawn from the evaluation pipeline. Under independence, the signal-to-noise ratio of the pipeline degrades multiplicatively across layers:

$$\mathrm{SNR}(\Pi) \leq \mathrm{SNR}(G) \cdot \mathrm{SNR}(S) \cdot \mathrm{SNR}(J)$$

Treating each V_i as the normalized SNR of the corresponding component then yields the bound. The motivating argument is that layer-wise validity losses multiply; this intuition is well-grounded even if the full formal derivation is not supplied here.

Remark 1 Because this is a conceptual bound rather than a proved proposition, equality in the inequality is not established. The bound becomes more informative (closer to the true V_{total}) when layer failures are less correlated, and less informative when positive correlation drives V_{total} below the product.

B.4 The Correlated-Failure Extension

Let $\epsilon_i = 1 - V_i$ denote the error at layer i, and let $\rho_{ij} = \mathrm{Corr}(\epsilon_i, \epsilon_j)$. When $\rho_{ij} > 0$ (positively correlated failures):

$$V_{\text{total}} < V_1 \cdot V_2 \cdot V_3 \qquad (\text{when } \rho_{ij} > 0)$$

The most common source of positive correlation: all three pipeline components use LLMs from the same provider family, causing systematic tendencies (verbosity preference, formatting bias, positional sensitivity) to manifest consistently across G, S, and J.

Mitigation. Cross-provider pipeline design reduces ρ_{ij} toward zero, making the independence bound more accurate and V_{total} higher.

B.5 Worked Numerical Example

Cross-provider design: $V_{\text{total}} \leq 0.70 \times 0.75 \times 0.75 = 0.394$

Within-family design with metric mismatch: $V_{\text{total}} \leq 0.70 \times 0.75 \times 0.58 = 0.305$

An evaluation score reported to two decimal places from a pipeline in the 0.30–0.40 validity range is precise to a degree of resolution the underlying measurement cannot support.

B.6 Estimation Protocol

For a pipeline operator who wants to estimate V_{total}:

1. **Estimate** V_1: generate 100 tasks from G and compare to 100 expert-curated reference tasks. Compute the proportion the reference judges rate as sampling the intended construct.
2. **Estimate** V_2: run a subset of tasks (minimum 50) with both the LLM simulator S and real human users; compute $\mathrm{ICC}(A, 1)$ between simulated and real agent success rates, separately for at least two user population groups.
3. **Estimate** V_3: use the appropriate IRR metric for the rubric's measurement scale (Figure 1).
4. **Compute the bound**: $V_{\text{total}} \leq V_1 \times V_2 \times V_3$. Report this bound alongside benchmark results.

If $V_{\text{total}} \leq 0.50$, benchmark scores should not be used for consequential deployment decisions without independent validation.