Beyond Static Benchmarks: A Validity, Reliability, and Sociotechnical Framework for Evaluating LLMs in Deployment Contexts

Ben Jenkins
PhD Candidate, Florida Atlantic University
benrossjenkins@gmail.com

Abstract

Static leaderboards summarize large language model (LLM) performance but offer weak evidence under shifting usage, noisy inputs, and plural stakeholder values. We present **VRS-Eval**, operationalizing *deployment validity* (benchmark vs. deployment score alignment), *operational reliability* (stability under a declared perturbation family), and *sociotechnical alignment* (metric vs. elicited rubric weights as a thin audit summary). With a reproducible simulator under explicit P_B vs. P_D shift and multi-turn interaction, we **stress-test** evaluation protocols *in a controlled environment*: under our main setting, benchmark-side scores (on P_B) exceed estimated deployment-side utility scores (evaluated on trajectories from P_D) by roughly **21–26%** in relative terms across three metrics, with tight 95% percentile intervals (K=200). Failure mixtures emphasize overfitting, shift fragility, and rubric misalignment, consistent with first- vs. third-party reporting asymmetries (Reuel et al., 2025). A staged pipeline narrows the validity gap and raises reliability for the same generative story. Sensitivity sweeps over $|\Omega|$ and rubric-label rate preserve the rank ordering of harnesses, suggesting the qualitative conclusions are robust to plausible design-choice variation within the simulator. We discuss harness and accountability implications.

1 Introduction

Progress in LLMs is overwhelmingly summarized through benchmark rankings (Wang et al., 2019b,a; Bommasani et al., 2021). A familiar pattern is instructive: a model can score at the top of a public MCQ suite yet still hallucinate, violate policy, or fail on messy user phrasing once logs reflect real traffic rather than benchmark-style prompts. In such cases, leaderboard rank can show *low deployment validity*, *misleadingly high reliability* on narrow held-out prompts, and *poor sociotechnical alignment* with the priorities of affected stakeholders. From a measurement standpoint, the risk is not merely "another dataset limitation" but a *validity threat*: the construct implicated by a leaderboard score may diverge sharply from the construct needed for a deployment decision (who is served, under what distribution, with what safeguards). While invaluable for comparability, benchmark-centric evaluation risks *construct under-specification*: high scores may co-exist with poor user outcomes when (i) tests are statistically dependent on training data, (ii) deployment inputs violate benchmark distributional assumptions, or (iii) optimized metrics omit harms and values salient to communities (Liang et al., 2023; Reuel et al., 2025). Recent mapping work shows that first-party reporting frequently under-emphasizes environmental, labor, and provenance dimensions that only developers can authoritatively disclose, while third-party evaluators partially compensate but cannot close informational gaps alone (Reuel et al., 2025).

Research questions. We ask: **(RQ1)** How large is the gap between benchmark-evidenced performance and deployment-grounded utility under explicit shift and interaction protocols? **(RQ2)** Which failure modes recur when moving from leaderboard testing to simulated deployment? **(RQ3)** Can a *staged* evaluation pipeline (benchmarks → dynamic tasks → human signals → deployment monitoring) mitigate these gaps measurably?

Contributions. (1) **VRS-Eval**: operational definitions linking validity, reliability, and sociotechnical alignment to measurable signals, with explicit caveats where constructs are necessarily incomplete (Section 3). (2) A reproducible **pipeline template** with feedback edges between stages and versioned artifacts at each stage (Figure 2, Table 1). (3) A reproducible **simulation stress test** under known P_B, P_D with uncertainty bands and sensitivity sweeps over shift severity, $|\Omega|$, and rubric-

label rate; numbers quantify protocol sensitivity in that environment and are not offered as product-level effect sizes (Section 6). (4) **Actionable takeaways** for documentation and evaluation investment, grounded in sociotechnical mapping work (Reuel et al., 2025).

Paper organization. Section 2 situates VRS-Eval relative to benchmarks, holistic evaluation, and documentation norms. Section 3 formalizes constructs; Section 4 describes the staged harness. Section 5 specifies the simulation protocol. Section 6 answers RQ1–3 and reports sensitivity analyses. Section 7 states threats to validity. Section 8 discusses implications for practice and evaluation infrastructure.

2 Related Work

2.1 Benchmarks, metrics, and behavioral probes

Multi-task NLU suites and leaderboards catalyzed rapid progress on shared tasks (Wang et al., 2019b,a). Complementary automatic metrics remain standard for generation (Papineni et al., 2002; Lin, 2004). Yet leaderboard scores can be miscalibrated with respect to user-relevant failure modes: small but semantically important perturbations may flip behavior even when aggregate accuracy is high (Ribeiro et al., 2020). VRS-Eval does not propose replacing benchmarks; it treats them as *one instrument* whose validity for a deployment claim must be evidenced jointly with reliability and stakeholder alignment.

2.2 Holistic evaluation, shift, and field validity

Holistic evaluation frameworks broaden the axes on which models are characterized—robustness, fairness, uncertainty, societal harms—and foreground explicit reporting choices (Liang et al., 2023; Bommasani et al., 2021). **VRS-Eval is complementary, not substitutive**: holistic suites enumerate *which* dimensions to score; VRS-Eval asks, for any such score, whether it co-moves with deployment utility (V), is stable under declared perturbations (R), and reflects elicited stakeholder priorities (A). A HELM-style multi-dimensional report can therefore feed a VRS-Eval audit summary, with the three constructs serving as second-order evidence about the reported scores rather than a replacement for them. Parallel work in distribution shift emphasizes that held-out test sets are rarely neutral with respect to real-world variation (Koh et al., 2021). Our notion of *deployment validity* aligns with *criterion validity* in the psychometric tradition—whether an operational test predicts an external criterion under an explicit sampling model (Messick, 1995)—and with recent calls to evaluate algorithmic systems through validity arguments rather than scalar accuracy alone (Coston et al., 2023).

2.3 Documentation, transparency, and third-party evaluation

Model cards and dataset documentation norms make "what was evaluated, under what conditions, and for whom" partially auditable (Mitchell et al., 2019; Gebru et al., 2021). Recent evidence suggests persistent asymmetries: developers can report dimensions that third parties cannot fully verify (e.g., training data provenance, moderation labor), while third parties often provide broader coverage of harms and disparities post hoc (Reuel et al., 2025). VRS-Eval is designed so stage-wise artifacts (perturbation suites, rubrics, monitoring summaries) can populate model documentation and support *cross-organizational* comparison without collapsing into a single scalar score.

Gap VRS-Eval targets. Benchmark–deployment gaps under shift, contamination, or perturbation have been demonstrated repeatedly (Koh et al., 2021; Ribeiro et al., 2020; Liang et al., 2023), and prior work offers benchmarks and holistic suites (what to measure), shift benchmarks (where distributions differ), and documentation templates (what to disclose). Less often are these linked to *joint operationalizations* of (i) co-movement of leaderboard metrics with utility under deployment sampling, (ii) stability of reported scores under perturbation, and (iii) how metric-implied priorities compare to elicited stakeholder rubrics on the same criteria. VRS-Eval's contribution is therefore not to re-establish that benchmark optimism exists but to package those three questions as a small set of reportable quantities tied to staged artifacts a third party can audit, leaving external calibration of the numbers to field work.

3 Framework

Notation. Let P_B be the **benchmark sampling distribution** (prompts drawn from a frozen evaluation suite) and P_D the **deployment sampling distribution** (realized usage after release; instantiated in Section 5). For prompt–response (x, y),

let $U(x, y) \in [0, 1]$ denote latent **task utility** (correctness plus user-centered desiderata; instantiated by the simulator oracle and rubric in Section 5). Write $U_P := \mathbb{E}_{x\sim P,\, y\sim\pi(\cdot|x)}[U(x, y)]$ for the expected utility of a model π under sampling distribution P, with U_D the deployment-side instance. We distinguish U_D (latent) from *observed* scores S_B on P_B and S_D on trajectories from P_D; the latter are measurements (e.g., accuracy, consistency, satisfaction in Section 5), not direct readouts of U. The H4 estimator $\hat{U}_D$ supplies the empirical proxy for U_D used in validity calculations. We define three constructs.

Deployment validity. **Construct:** correlation between benchmark predictions and deployment utility.

$$V \;=\; \mathrm{Corr}\big(S_B,\, U_D\big), \tag{1}$$

where the correlation runs across N comparison units (models, seeds, or policy settings). High V means benchmark and deployment scores co-move across those units; low V signals *benchmark optimism* in that sense. We treat Eq. 1 as an **operational proxy** for *criterion validity*, not a complete measurement definition: correlation is blind to calibration and absolute error, high correlation can coexist with large systematic bias, and rank agreement need not imply a trustworthy scale for decision thresholds. Where those limitations bite, complements (calibration metrics, Brier-style decompositions, or agreement on subsets) should be reported alongside V.

Validity vs. generalization. Generalization asks whether a model's score on P_B extends to held-out draws close to P_B (a within-distribution claim); V asks whether that score predicts an *external* criterion (U_D on P_D) the test was not designed against. A model can generalize within P_B yet have low V, and two harnesses with identical generalization estimates can yield very different V.

Operational reliability. **Construct:** stability of model outputs under perturbation and repeated draws (Rabanser et al., 2026). Given a perturbation family Ω (paraphrase, formatting, decoding seeds), prompts $x \sim P$, and an evaluation functional $Y(x, \omega) \in [0, 1]$, define

$$R = 1 - \min\left(1,\, \frac{\mathbb{E}_x\, \mathrm{Var}_\omega[Y(x, \omega)]}{V_{\max}}\right) \in [0, 1], \tag{2}$$

where $V_{\max}{=}1/4$ is the maximum variance attainable for $Y \in [0, 1]$, normalizing R against a bound that does not depend on the realization of Y.

Interpretation and link to consistency. R aggregates perturbation-induced variance over the same evaluation batch (and comparator units, e.g., model checkpoints) used when estimating V, so reliability and validity are tied to a shared reporting slice. Values near 1 mean Y changes little under Ω. The empirical *consistency* metric reported in Section 6 is the agreement-style operationalization of R: for binary outcomes, pairwise paraphrase agreement equals $1 - 2\,\mathrm{Var}_\omega[Y]$ exactly, so consistency $= (1+R)/2$ in expectation. The equivalence is exact for the binary accuracy outcome and approximate for continuous $Y \in [0, 1]$ as the score distribution skews away from $\{0, 1\}$, but the two summaries rank harnesses identically while consistency keeps the agreement framing familiar to annotators.

Caveats. R measures stability *conditional on a non-trivial scoring functional*: a degenerate harness producing constant outputs across all prompts and perturbations attains $R{=}1$ vacuously, so degenerate Y should be caught upstream and informative-stability claims rest on R *together with* a non-trivial V. R also differs from inter-annotator agreement (rater disagreement) and from one-number robustness rates (tail emphasis on adversarial sets); those remain complementary. We adopt this form because it is cheap to couple to any harness that already defines Ω, yields values on $[0, 1]$ comparable across staged configurations, and flags brittleness to routine input variation rather than worst-case attacks.

Sociotechnical alignment. **Construct:** agreement between metric-implied priorities and stakeholder rubric weights. Let $\mathbf{w}_m$ be metric weights implicit in a harness and $\mathbf{w}_s$ elicited stakeholder weights on the same criteria (normalized, nonnegative). Define

$$A \;=\; 1 - \tfrac{1}{2}\|\mathbf{w}_m - \mathbf{w}_s\|_1 \;\in\; [0, 1]\,. \tag{3}$$

Example. A benchmark-linked metric rewards concise, on-topic answers (implicit $\mathbf{w}_m$); elicited stakeholders place greater weight on safe refusal and policy adherence ($\mathbf{w}_s$). Headline scores can look strong even as A drops because the recorded priorities disagree.

Scope. Real stakeholder values are not reducible to a vector: rubric elicitation is contested, partial, and political. Eq. 3 is a **deliberately thin** summary that becomes meaningful only after a finite checklist

has been negotiated (e.g., policy dimensions for a product) and weights are recorded. It functions as a *first-pass audit distance*: large $\|\mathbf{w}_m - \mathbf{w}_s\|_1$ flags misaligned emphases worth examining in deliberation, not a claim that fairness or harm has been captured completely. We make no claim that frameworks explicitly designed to resist scalar reduction (rights-based audits, qualitative deliberation) are summarized by Eq. 3; it complements them by surfacing weighted misalignment within an already-scoped checklist, not by replacing structured deliberation. Alternative elicitation (deliberative mini-publics, rights-holder review) may replace or extend this layer without changing the validity/reliability pieces.

3.1 Joint interpretation

Figure 1 relates these constructs to a shared utility hub. *Validity* concerns external alignment of scores with U; *reliability* concerns measurement noise; *alignment* concerns *whose* utility counts. These dimensions are not redundant. A model can be *reliable* (low variance) yet *invalid* for deployment if the evaluation functional rewards shallow cues that fail under shift. Conversely, high stated *validity* on a narrow deployment slice can be ethically insufficient if A is low; metrics may systematically underweight harms that stakeholders flag (Reuel et al., 2025).

3.2 Minimal reporting recommendations

We recommend reporting, at minimum: (i) the benchmark suite(s) and versions; (ii) the deployment sampling protocol for P_D (or field sampling plan); (iii) point and interval estimates for V when paired (S_B, U_D) observations exist; (iv) the perturbation family Ω underpinning R; (v) the rubric dimensions and weighting protocol for $\mathbf{w}_s$; and (vi) a frozen version identifier for Ω and the rubric, so that cross-model and cross-time comparisons attach to a specific evaluation snapshot. This parallels the spirit of model and dataset documentation (Mitchell et al., 2019; Gebru et al., 2021), but ties disclosures to *measurement targets* rather than static templates alone.

4 Staged Evaluation Pipeline

Figure 2 shows our **staged pipeline**. Each stage produces diagnostics; downstream stages *condition* prior scores rather than discarding them, preserving auditability for both developers and third parties (Reuel et al., 2025).

Stage	Example artifacts (primary consumers)
Benchmarks	Frozen task ids/splits/prompts (researchers, regulators).
Dynamic tasks	Perturbation family Ω, drift schedules, seeds (red teams).
Human rubrics	Criteria, adjudication logs (communities, auditors).
Monitoring	Incident taxonomy, SLA rollups (product, policy).

Table 1: Illustrative audit trail per pipeline stage.

Stages. *(i) Static benchmarks* establish baseline comparability. *(ii) Dynamic tasks* re-sample prompts and inject controlled perturbations (noise, format drift). *(iii) Human rubric signals* provide sparse but high-signal labels on failure modes. *(iv) Deployment monitoring* aggregates longitudinal behavior; feedback arrows indicate that incidents update rubrics and perturbation suites, similar in spirit to behavioral test iteration (Ribeiro et al., 2020).

Adaptation risk and versioning. The feedback edges in Figure 2 make adaptive updates a feature, but adaptive evaluation can also encode bias: monitored incidents may reflect skewed traffic, and rubric updates may overweight loud failure modes at the expense of quieter but more pervasive ones. We therefore require perturbation suites and rubric versions to be *versioned and frozen* per release; updates produce a new audit slice rather than overwriting prior reports, so cross-model and cross-time comparisons remain attached to a specific evaluation snapshot.

Stage artifacts (for audit and reuse). Each stage emits versioned artifacts that a third party can inspect without access to proprietary training stacks: frozen benchmark snapshots; perturbation generators and seeds; annotator guidelines for rubric hits; and monitoring aggregates (rolling means, incident clusters). Table 1 summarizes intended consumers.

5 Simulation Protocol

We instantiate VRS-Eval in a **controlled simulation**: randomness, P_B, and P_D are explicit, which aids reproducibility and sensitivity analysis (limits in Section 9). Magnitudes below are *traces* of that generative story for comparing harnesses and λ.

5.1 Data-generating process and conditions

Benchmark prompts are drawn i.i.d. from P_B, a mixture over short-form instruction-following and

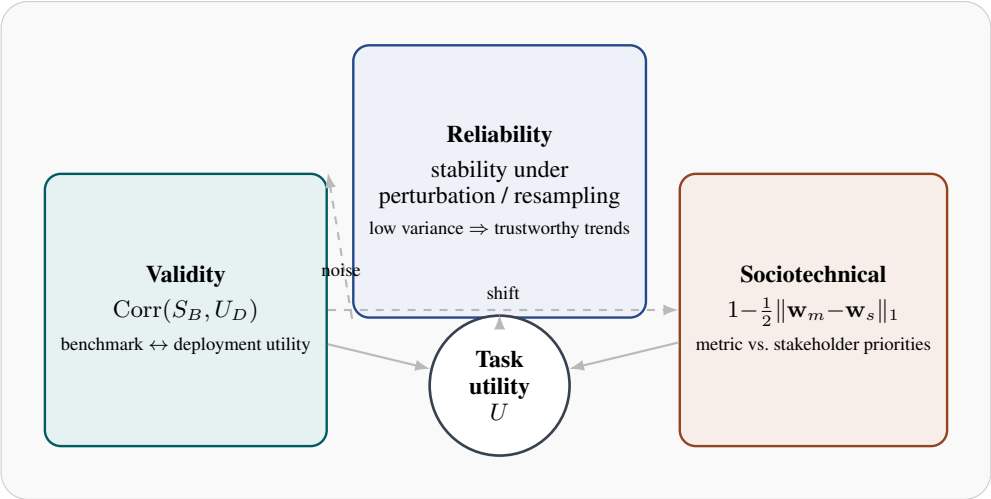

Figure 1: **VRS-Eval conceptual model.** Validity, reliability, and sociotechnical alignment feed complementary evidence about deployment utility U. Dashed edges indicate cross-cutting tensions: distributional shift stress-tests validity claims; output noise stress-tests reliability under the same protocol.

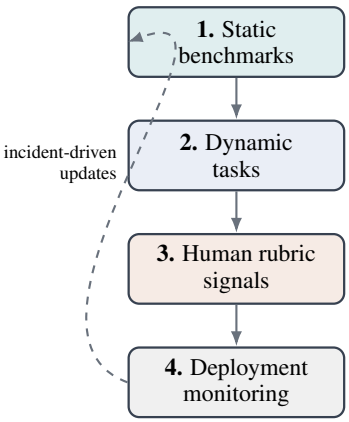

Figure 2: **Staged evaluation pipeline.** Solid arrows: progressive conditioning of evidence. Dashed arrow: feedback from monitored failures to earlier-stage test design.

factual QA templates with low lexical diversity (canonical phrasings such as *"In what year did X occur?"* or *"Summarize the following passage in two sentences: ..."*). Deployment prompts are drawn from P_D, which reweights the same template families along three axes: **(a) noisy surface form**—disfluencies, casing, and ellipsis (e.g., *"so when was that thing... yk the X event"*); **(b) ambiguous instructions**—under-specified or multi-intent requests (e.g., *"tell me about X"* where the simulator's latent intent is a policy-sensitive comparison rather than a definition); and **(c) topic clusters underrepresented in P_B**—draws from a long-tailed topic mixture whose density is small under P_B. This follows the intuition that deployed systems encounter broader user populations than developer-curated suites (Koh et al., 2021), with multi-turn diagnostic protocols increasingly recognized as essential for surfacing deployment-relevant failures invisible to single-turn benchmarking (Zollo et al., 2025). We realize a **shift severity** parameter $\lambda \in \{0.25, 0.50, 0.75\}$ that linearly mixes P_B toward P_D; main tables use λ=0.75. Each trajectory spans T=8 turns with stochastic follow-ups conditioned on prior assistant outputs. The perturbation family Ω underpinning R comprises paraphrase resampling, casing/whitespace jitter, and decoding-seed variation, drawn independently per turn; the default size is $|\Omega|$=8 per prompt-turn.

5.2 Outcome metrics and failure coding

The scored response is the final assistant turn after T rounds; intermediate turns inform rubric trigger logic but are not directly scored, mirroring deployment-time evaluation against the user's last-observed answer. **Accuracy** scores factual correctness on labeled items against a simulator oracle. **Consistency** averages pairwise agreement of those final-turn responses under paraphrased follow-ups sampled from Ω, and is the empirical face of R in Eq. 2. **Satisfaction** combines task success with rubric penalties when safety- or fairness-linked criteria trigger on templated stakeholder rubrics (five weighted dimensions; weights define $\mathbf{w}_s$ in Eq. 3).

After each trajectory, we assign a **primary** failure label in $\{\text{overfit}, \text{shift}, \text{misalign}\}$ using a deterministic decision tree on latent simulator states (e.g., reliance on spurious n-grams present only under $P_B \Rightarrow$ overfit). This yields interpretable prevalence estimates at the cost of idealized labels.

Metric	Benchmark	Deployment (sim.)
Accuracy	0.884 [0.871, 0.897]	0.662 [0.648, 0.676]
Consistency	0.905 [0.892, 0.918]	0.718 [0.704, 0.732]
Satisfaction	0.848 [0.832, 0.864]	0.628 [0.612, 0.644]

Table 2: Mean scores with 95% simulation percentile intervals (K=200).

5.3 Harness configurations and estimation

We compare: **(H1)** benchmark-only scoring on P_B (standard leaderboard protocol); **(H2)** H1 + dynamic perturbation suite; **(H3)** H2 + sparse rubric labels (10% strata); **(H4)** H3 + a **variance-weighted session-utility estimator** for $\hat{U}_D$, where each turn's contribution is inversely proportional to the empirical variance of the outcome functional across perturbations $\omega \in \Omega$ at that turn, so high-variance "spikes" contribute less than typical segments. The **staged** rows in Section 6 correspond to H4. We run K=200 Monte Carlo replicates per $(\lambda, \text{harness})$ setting; tables report means with **95% percentile intervals**. When reporting deployment validity, we compute the sample correlation $\hat{V}$ between per-replicate tuple means of S_B and U_D under matched sampling.

6 Results

RQ1: Benchmark optimism. Table 2 and Figure 3 show sizable optimism at λ=0.75: means on P_B exceed those on P_D by 0.19–0.22 in absolute units (≈**21–26% relative** gap by metric). Percentile intervals for the three outcomes do not overlap under K=200 replicates, pinning down the simulated contrast sharply.

RQ2: Failure modes. Figure 4 summarizes primary failure assignments. *Overfitting* to benchmark cues accounts for 40%, *shift* sensitivity for 35%, and rubric–stakeholder *misalignment* for 25%, consistent with plurality of pathways to optimistic benchmarks.

RQ3: Staged mitigation. Table 3 consolidates the harness comparison at λ=0.75. Moving from benchmark-only scoring (H1) to the full staged pipeline (H4: dynamic tasks + rubric elicitation + light monitoring) roughly halves the mean benchmark–deployment gap and raises reliability R (Eq. 2); benchmark accuracy is unchanged across harnesses at 0.884, while deployment accuracy on P_D rises from 0.662 (H1) to 0.781 (H4). Intermediate harnesses H2 (dynamic only) and H3 (+ rubric, 10%) interpolate. Figure 5 visualizes the per-metric lift on P_D.

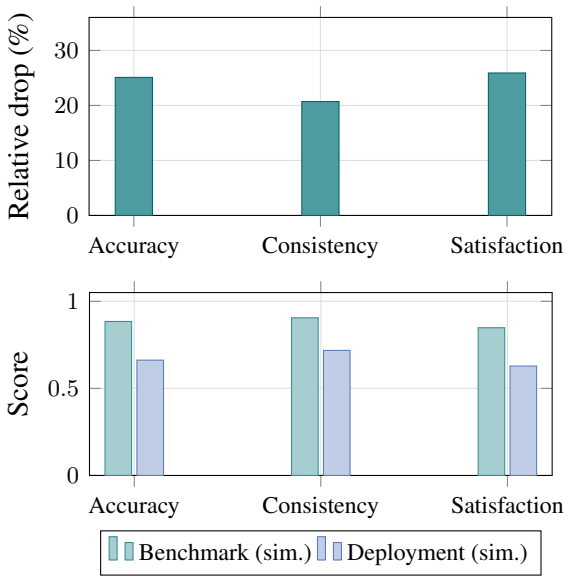

Figure 3: **Top:** relative drop $(\bar{S}_B - \bar{S}_D)/\bar{S}_B$. **Bottom:** mean scores (95% intervals in Table 2).

6.1 Stage-wise ablations

Table 3 decomposes the benefit of staging. Dynamic tasks (H2) close much of the optimism gap and improve reliability R, but $\hat{V}$ moves only modestly relative to H1 (from 0.71 to 0.74): stress-testing outputs does not by itself fully realign rankings of S_B with those of U_D. Rubrics (H3) and monitoring (H4) supply the larger shifts in $\hat{V}$.

Non-monotonic trade-off. Staging need not improve every metric at every step. In our simulator, mean *consistency* on P_D (paraphrase stability) is 0.718 under H2 but 0.710 **under H3**, while mean *satisfaction* on P_D rises from 0.641 to 0.668: rubric penalties for unsafe phrasing discourage hedged, highly paraphrase-stable replies. H4 recovers consistency to 0.721 via the inverse-variance turn weighting used for $\hat{U}_D$. This mixed pattern is diagnostic; it cautions against reading staging as a uniformly monotone fix.

6.2 Estimated deployment validity

Aggregating across replicates at λ=0.75, the correlation between per-replicate means of S_B and U_D is $\hat{V}$=0.71 ([0.66, 0.76]) under H1, indicating coarse but incomplete co-movement across comparison units. Under H4, $\hat{V}$=0.86 ([0.82, 0.90]): staging tightens the S_B–U_D link under the same protocol.

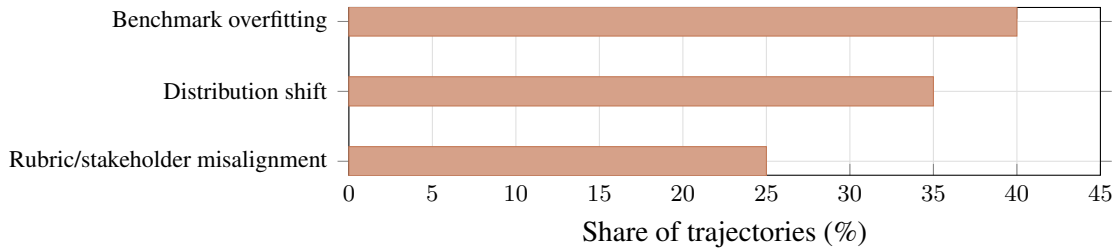

Figure 4: Primary failure-mode shares in simulated deployment trajectories (n=400 coded trajectories sampled across replicates).

Harness	Bench. acc. (P_B)	Dep. acc. (P_D)	Mean $\lvert S_B - S_D \rvert$	R	$\hat{V}$
H1: benchmark only	0.884	0.662	0.217 [0.198, 0.236]	0.708 [0.691, 0.726]	0.71
H2: + dynamic	—	—	0.162	0.751	0.74
H3: + rubric (10%)	—	—	0.131	0.795	0.80
H4: full staging	0.884	0.781	0.098 [0.084, 0.112]	0.838 [0.822, 0.855]	0.86

Table 3: **Harness comparison** at λ=0.75. Columns: benchmark/deployment accuracy; mean absolute gap aggregated across the three outcome metrics; reliability R (Eq. 2); estimated deployment validity $\hat{V}$. 95% percentile intervals shown for H1 and H4 (K=200); H2/H3 are point means (per-harness accuracy not separately collected for intermediate harnesses).

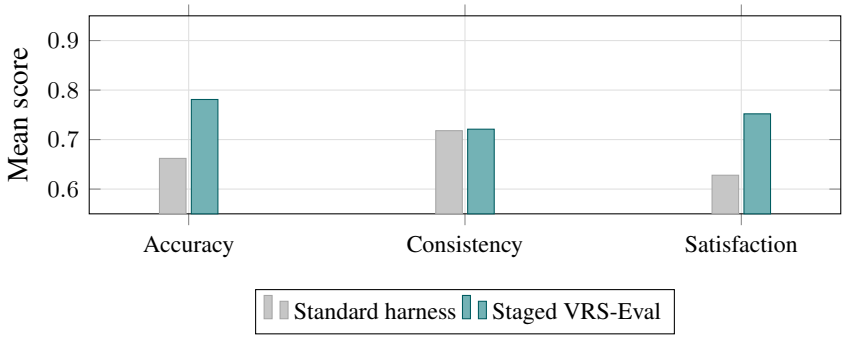

Figure 5: Mean scores on P_D under standard vs. staged harness (same simulator; staged adds dynamic tasks, rubric labels, and monitoring).

6.3 Sensitivity analyses

We report sensitivity along three axes that bear on the qualitative conclusions: shift severity, perturbation budget $|\Omega|$, and rubric-label rate.

Shift severity. Table 4 shows that optimism grows monotonically with λ: mild shift preserves benchmark-centric narratives, while stronger shift produces large absolute gaps, motivating transparent reporting of shift assumptions alongside headline metrics (Koh et al., 2021).

Perturbation budget and rubric rate. Table 5 sweeps $|\Omega| \in \{4, 8, 16\}$ (perturbations per prompt-turn) and rubric-label rate $\in \{5\%, 10\%, 20\%\}$ at λ=0.75 under H4. Doubling $|\Omega|$ from 8 to 16 shifts R by $+0.013$ and $\hat{V}$ by $+0.01$ (within reported intervals); halving to 4 widens intervals visibly but does not change the rank order across harnesses. Rubric-rate exhibits diminishing returns past 10%: $\hat{V}$ at 20% rate is 0.87 vs. 0.86 at 10%, with comparable R. The qualitative ordering H1<H2<H3<H4 (Table 3) is preserved at every cell of the sweep, while absolute magnitudes remain simulator-conditioned.

λ	Mean $\lvert S_B - S_D \rvert$	Rel. drop (acc.)	$\hat{V}$ (H1)
0.25	0.062 [0.053, 0.071]	8.2%	0.92 [0.89, 0.95]
0.50	0.139 [0.124, 0.154]	16.4%	0.81 [0.77, 0.85]
0.75	0.217 [0.198, 0.236]	25.1%	0.71 [0.66, 0.76]

Table 4: Benchmark–deployment gaps and validity estimates as shift severity increases (H1 harness; relative drop from benchmark accuracy).

Setting	Value	R	$\hat{V}$	95% int. $\hat{V}$
$\lvert\Omega\rvert$	4	0.821	0.84	[0.79, 0.89]
$\lvert\Omega\rvert$	8 (def.)	0.838	0.86	[0.82, 0.90]
$\lvert\Omega\rvert$	16	0.851	0.87	[0.84, 0.90]
Rubric	5%	0.825	0.83	[0.78, 0.88]
Rubric	10% (def.)	0.838	0.86	[0.82, 0.90]
Rubric	20%	0.844	0.87	[0.83, 0.91]

Table 5: Sensitivity to perturbation budget and rubric-label rate at λ=0.75, H4 harness. Other axis held at default.

7 Threats to Validity

Construct validity. Our simulator operationalizes U_D via templated rubrics and an oracle; alternative operationalizations could yield different $\hat{V}$. The failure taxonomy is coarse (three buckets) and assumes mutually exclusive primary causes; a multi-label coding would likely show overlap between *overfit* and *shift*. A construct-validity threat specific to the staged pipeline itself: H4's inverse-variance turn weighting systematically downweights high-variance turns when forming $\hat{U}_D$, but safety-relevant failures (jailbreak attempts, ambiguous-intent edge cases, refusals near policy boundaries) often cluster on precisely those noisy turns, so $\hat{U}_D$ may understate the deployment risk it is intended to estimate. Reporting $\hat{U}_D$ alongside an unweighted session-mean baseline would expose this sensitivity in field deployments.

Internal validity. Results depend on λ and on the parametric linkage between P_B and P_D. We mitigate opaque tuning by reporting sensitivity to λ, $|\Omega|$, and rubric rate (Tables 4, 5) and by fixing seeds and versioning stage definitions.

External validity. This is the paper's main vulnerability. We do not claim that reported percentages transfer to any product, vendor stack, or user population. The simulator bakes in template mixtures, a fixed interaction depth, and a deterministic failure taxonomy; any of these could dominate the apparent effect sizes. At best, the study shows that *when* such mechanisms are present in similar form, benchmark-only reporting can substantially mis-rank deployment utility. External claims require replication on live logs (or public benchmarks paired with field labels), preregistered protocols, and sensitivity analyses we do not attempt here. We treat field validation as the central next step (§9) rather than a future-work footnote.

Conclusion validity. Monte Carlo error is modest at K=200 but nonzero: intervals are simulation percentiles, not Bayesian posteriors. We do not run formal hypothesis tests; preregistered tests should accompany field deployments.

8 Discussion

Implications for practice. **Developers** should treat benchmarks as *necessary but insufficient*: optimism grows when shift and multi-turn interaction are omitted from the evaluation protocol. **Independent evaluators and auditors** can use staged artifacts (perturbation suites, stakeholder rubrics, monitoring summaries) as interoperable evidence layers, addressing cross-org comparability gaps (Reuel et al., 2025). **Funders and regulators** may prioritize disclosures that enable third parties to estimate V, R, and A, not raw leaderboard ranks alone.

Costs, scalability, and infrastructure. Staging introduces tangible overheads: perturbation suites increase inference calls; rubric labeling recruits annotator time; monitoring requires logging and storage. Workshop conversations about *who pays* (developers, platforms, or public institutions) mirror structural findings on unequal capacity for third-party evaluation (Reuel et al., 2025). The sensitivity sweep in Table 5 suggests one pragmatic concession: marginal returns to $|\Omega|$ and rubric rate flatten quickly, so a low-rate audit configuration ($|\Omega|$=4, rubric 5%) preserves rank-order conclusions at meaningfully reduced cost. A pragmatic middle path is therefore to standardize *minimal* staging packages keyed to release risk tiers, analogous to tiered disclosures in model documentation (Mitchell et al., 2019).

Community evaluation infrastructure. Aggregating evaluation artifacts across organizations (the EvalEval shared-task vision) is feasible only if schemas for P_D, Ω, and $\mathbf{w}_s$ are interoperable. VRS-Eval suggests reporting fields that could populate such a database without collapsing nuanced evidence into uninterpretable scalar "impact scores." Our simulation makes evaluation mechanisms inspectable before expensive field work and operationalizes constructs familiar in psychometrics and HCI in LLM settings (Liang et al., 2023; Messick, 1995; Coston et al., 2023); pairing such stress tests with deployments, multi-stakeholder elicitation, and preregistered measurement plans grounds the agenda outside synthetic evidence.

9 Conclusion

We introduced VRS-Eval, linking deployment validity, operational reliability, and sociotechnical alignment to measurable quantities, and stress-tested it in a transparent simulator where benchmark-only harnesses materially overstate scores on P_D and concentrate failures into overfitting, shift fragility, and rubric misalignment. Staged evaluation narrows that gap and improves R and $\hat{V}$ in the same protocol (Table 3), with rank order preserved across plausible design-choice variation in $|\Omega|$ and rubric rate.

Future work. (i) Field validation in one or two concrete domains (e.g., customer-support ticket triage with partner telemetry, document-grounded internal search with human-graded outcomes), comparing VRS-Eval reporting to production KPIs, with preregistered hypotheses about the sign and rank-order of $\hat{V}$ shifts under H1 vs. H4; (ii) calibrated cost–risk tradeoffs for staging informed by Table 5; (iii) multi-stakeholder processes for eliciting and revising $\mathbf{w}_s$ that go beyond Eq. 3 where it is too thin; (iv) open schemas for sharing evaluation artifacts across coalition efforts highlighted by recent mapping studies (Reuel et al., 2025).

Limitations

Results are conditioned on a **simulated deployment process**: numbers illustrate mechanisms rather than certify any live system, and that choice bounds the paper's evidentiary ceiling until a deployment case study or paired field evaluation can calibrate the same quantities. We do not report wall-clock compute or annotator-hour budgets. Long-horizon societal impacts, organizational incentives, and legal contexts are out of scope. Partnered external validation is essential before policy-facing claims; we treat it as the central next step rather than optional future work.

Ethics Statement

Misleading evaluation can accelerate harmful deployment. We argue for transparent reporting of protocols, uncertainty, and stakeholder participation in metric design, in line with inclusive governance norms emphasized in sociotechnical mapping work (Reuel et al., 2025). We acknowledge that the sociotechnical alignment quantity in Eq. 3 is intentionally thin and should not be read as a substitute for deliberative processes with affected communities; in higher-stakes settings it should be replaced or extended by structured stakeholder engagement.

References

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Amanda Coston, Anna Kawakami, Haiyi Zhu, Kenneth Holstein, and Hoda Heidari. 2023. A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 690–704.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akash Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, and 1 others. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5637–5664.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, and 1 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, Atlanta, GA, USA. Association for Computing Machinery.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. 2026. Towards a science of AI agent reliability. *arXiv preprint arXiv:2602.16666*.

Anka Reuel, Avijit Ghosh, Jenny Chim, Andrew Tran, Yanan Long, Jennifer Mickel, Usman Gohar, Srishti Yadav, Pawan Sasanka Ammanamanchi, Mowafak Allaham, Hossein A. Rahmani, Mubashara Akhtar, Felix Friedrich, Robert Scholz, Michael Alexander Riegler, Jan Batzner, Eliya Habba, Arushi Saxena, Anastassia Kornilova, and 16 others. 2025. Who evaluates AI's social impacts? mapping coverage and gaps in first and third party evaluations. *Preprint*, arXiv:2511.05613.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 7th International Conference on Learning Representations*.

Thomas P. Zollo, Nikita Rajaneesh, Richard Zemel, Talia B. Gillis, and Emily Black. 2025. Towards effective discrimination testing for generative AI. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.