DID THE MODEL GET BETTER OR JUST DIFFERENT? DIRECTIONAL EVALUATION OF LLM UPDATES FROM UNLABELED PROBES

Anonymous authors
Paper under double-blind review

ABSTRACT

Model updates can change behavior without revealing whether task reliability improved or degraded. We show that this distinction is structural: symmetric behavioral distances can quantify change magnitude but cannot consistently encode the sign of a reliability change. We introduce the Directional Disagreement Estimator (DDE), a parameter-free estimator that combines the fraction of unlabeled probes on which two model versions disagree with a blind judge's net preference over those disagreements. Under a simple judge-error model, DDE is attenuated toward zero as judge discriminative accuracy approaches chance. In a pre-specified confirmatory evaluation on GPQA across 22 model-update transitions, our primary endpoint — the correlation of DDE with held-out signed accuracy change — is Spearman $\rho = 0.600$ (95% CI $[0.22, 0.87]$, $p = 0.0031$), and DDE attains AUROC 0.85 for detecting regressions exceeding five accuracy points. A secondary sign-accuracy comparison gives 0.68 against the 0.50 constant-sign baseline on a balanced set. DDE does not outperform a zero-change predictor on absolute error, so we position it as a directional statistic rather than a calibrated estimate of magnitude. Symmetric answer disagreement instead predicts change magnitude but not direction. A judge measured above chance yields useful DDE signal; one measured at chance yields none. Labeled canaries remain stronger, positioning DDE as a screening method that needs no probe labels rather than a replacement for labeled evaluation.

1 INTRODUCTION

Large language models deployed through APIs are not static systems. Providers replace model snapshots, modify inference infrastructure, change decoding behavior, and update prompts or scaffolds. Even when an endpoint remains operational, its behavior can shift substantially. Recent work has therefore proposed black-box behavioral fingerprints that detect changes in an endpoint's output distribution without requiring access to model weights or provider internals (Leshin et al., 2026). Such monitoring can answer an operationally important question: *did the model change?*

For many applications, however, detecting change is not enough. A deployment team deciding whether to accept, investigate, or roll back an update needs to know whether task reliability became *better or worse*. These are different questions. A behavioral distance may increase because a new model corrects previous mistakes, because it introduces new mistakes, or because it changes style while preserving accuracy. Similarly, statistical degradation testing can determine whether a measured drop in labeled benchmark accuracy is likely to reflect a genuine regression rather than finite-sample noise, but it presupposes access to correctness labels (Kübler et al., 2026).

We study the intermediate setting: an old and a new model version can be queried on a fixed probe set, but ground-truth labels for those probes are unavailable or expensive. Can their responses nevertheless reveal the direction of the reliability change?

We begin with a simple structural observation. Most behavioral-change measures are symmetric: exchanging the old and new versions leaves their value unchanged, $D(M_{\text{old}}, M_{\text{new}}) = D(M_{\text{new}}, M_{\text{old}})$. Reliability change is instead directional. If $\Delta p = p_{\text{new}} - p_{\text{old}}$, then reversing the transition changes its sign. Consequently, a symmetric statistic alone cannot encode the orientation of a non-zero

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reliability change. It may tell us how much behavior changed, but not whether the change was an improvement or a regression.

This distinction is practically important rather than merely notational. In the pre-specified GPQA evaluation, the symmetric agreement distance is nearly unrelated to signed accuracy change ($\rho = -0.060$) but substantially associated with its magnitude ($\rho = 0.565$), while a blind pairwise judge shows the opposite pattern. Reverse-transition diagnostics make the constraint explicit: the agreement distance is bit-identical after reversing old and new, whereas the target Δp changes sign.

Motivated by this separation, we define a simple *Directional Disagreement Estimator* (DDE). Given N unlabeled probes, we first identify the subset on which the two model versions produce different answers. A blind judge, given neither model identity nor correctness labels, compares the old and new response on each such item. If W_{new} and W_{old} denote the number of judge preferences for the new and old responses, respectively, then

$$\widehat{\Delta p}_{\text{DDE}} = \frac{W_{\text{new}} - W_{\text{old}}}{N}. \tag{1}$$

Equivalently, DDE is the product of the parsed-answer disagreement rate d_{parsed}, which supplies a notion of scale, and the judge's net preference J, which supplies orientation. The estimator contains no fitted coefficient and requires no probe labels.

DDE is not expected to work with an arbitrary judge. Under a simple symmetric judge-error model, if the judge chooses the actually correct response with probability q on items where exactly one model is correct, then $\mathbb{E}[\widehat{\Delta p}_{\text{DDE}}] \approx (2q-1)\Delta p$. The estimator therefore attenuates toward zero as judge accuracy approaches chance and should carry no directional information when $q = 0.5$. This gives the method a measurable failure condition rather than assuming that any sufficiently large LLM is a reliable evaluator.

We evaluate these claims in a frozen confirmatory experiment on GPQA, a 448-question expert-written benchmark in biology, physics, and chemistry (Rein et al., 2024). Fourteen model versions answer all benchmark items; 64 questions are frozen as unlabeled probes and 384 disjoint questions are used only to estimate the held-out reliability change across 22 primary transitions. DDE is specified before this confirmatory evaluation and is not tuned on a development set.

Our main findings are threefold. First, DDE is associated with signed held-out reliability change at Spearman $\rho = 0.600$ (95% CI $[0.22, 0.87]$, $p = 0.0031$), with 0.68 sign accuracy and AUROC 0.85 for detecting regressions exceeding five accuracy points. Second, the direction–magnitude separation predicted above appears empirically (Table 1). Third, judge quality is consequential: on the same 16 model-replacement transitions a GPT-5 judge with measured accuracy $q = 0.657$ yields $\rho = 0.651$, while a GPT-4o judge at chance ($q = 0.499$) yields none ($\rho = -0.140$).

Labels remain substantially stronger when available: thirty-two labeled canary items reach $\rho = 0.847$ and AUROC 0.98. We therefore view DDE as a screening mechanism for settings where probe labels are unavailable or costly, not a replacement for conventional labeled evaluation.

Our contributions are: (i) *a directional measurement constraint* — symmetric behavioral-change statistics cannot by themselves encode the sign of a reliability change, demonstrated with reverse-transition diagnostics; (ii) *DDE*, a parameter-free screening score needing no probe labels, which beats every constant-sign baseline but not a zero-change predictor on absolute error; and (iii) *a measurable judge-quality condition* derived from the estimator itself, under which we observe both success and failure.

2 Problem Formulation

2.1 Signed reliability change

Let M_o denote an old model version and M_n an updated version. For an input $x \sim \mathcal{D}$, let $C_o(x), C_n(x) \in \{0, 1\}$ denote whether the responses of M_o and M_n are correct under the target task, and let $p_o = \mathbb{E}_x[C_o(x)]$ and $p_n = \mathbb{E}_x[C_n(x)]$ be the corresponding reliabilities. The quantity of interest is the signed reliability change

$$\Delta p = p_n - p_o, \tag{2}$$

positive for an improvement and negative for a regression.

During monitoring we assume access to a probe set $\mathcal{P} = \{x_1, \ldots, x_N\}$ and to responses from both versions, but not to the ground-truth labels for these probes. The objective is therefore not to estimate conventional benchmark accuracy directly. Instead, we ask whether information in the paired, unlabeled responses can estimate the orientation and approximate scale of Δp.

2.2 Symmetric statistics cannot determine direction

Consider any statistic $S(M_o, M_n)$ that is symmetric with respect to the two model versions, $S(M_o, M_n) = S(M_n, M_o)$. Answer-disagreement rates and standard symmetric distances between response distributions are of this form. By contrast, Δp is antisymmetric: $\Delta p(M_n, M_o) = -\Delta p(M_o, M_n)$.

Proposition 1 (Symmetric statistics cannot encode orientation)**.** *For any symmetric statistic S, no deterministic function $f(S)$ can equal the signed reliability change for both orientations of every pair with non-zero Δp.*

Proof. Because $S(M_o, M_n) = S(M_n, M_o)$, any deterministic f satisfies $f(S(M_o, M_n)) = f(S(M_n, M_o))$: the value is unchanged under reversal. But for a transition with $\Delta p \neq 0$ we have $\Delta p(M_o, M_n) = -\Delta p(M_n, M_o)$, so the two targets differ. A single value of $f(S)$ cannot equal both. □

This does not imply that symmetric statistics are useless. They may contain substantial information about $|\Delta p|$, or more generally about the extent of behavioral change; the constraint is specifically about orientation. Our reverse-transition diagnostics instantiate the argument directly on the evaluation pipeline used in our experiments (Section 5.3). Appendix B gives an expanded discussion of the constraint and its scope.

3 Directional Disagreement Estimation

3.1 Paired correctness and response disagreement

Suppose labels were available temporarily for analysis. Partitioning the N probe examples by paired correctness, let $b = \#\{x : C_o(x) = 1, C_n(x) = 0\}$ be the items the update breaks and $c = \#\{x : C_o(x) = 0, C_n(x) = 1\}$ the items it fixes. The empirical accuracy difference is then exactly

$$\Delta p = \frac{c - b}{N}, \tag{3}$$

since items on which both versions are correct, or both incorrect, contribute equally to the two accuracies and cancel. This paired decomposition is closely related to the discordant-pair perspective used in statistical degradation testing (Kübler et al., 2026), but our setting differs in one crucial respect: during estimation, the correctness variables are hidden.

What is observable is which answers changed. Writing $a_o(x)$ and $a_n(x)$ for the parsed answers, define the response-disagreement set $\mathcal{D} = \{x \in \mathcal{P} : a_o(x) \neq a_n(x)\}$, with $m = |\mathcal{D}|$ and disagreement rate

$$d_{\text{parsed}} = \frac{m}{N} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[a_o(x_i) \neq a_n(x_i)]. \tag{4}$$

Two unparseable responses count as *agreeing* here, since neither yields a label to disagree on. This is the gating rate DDE uses; it is not the agreement-distance baseline $d_{\text{agreement}}$ defined in Section 4. For single-answer multiple choice, correctness can only change inside $\mathcal{D}$: if both versions emit the same label they are both right or both wrong. But $\mathcal{D}$ also contains items where both models are wrong and merely choose different incorrect answers. Writing e for the number of such both-wrong disagreements,

$$m = b + c + e. \tag{5}$$

The gating rate d_{parsed} therefore quantifies how often the update changes the observable answer, but cannot say whether those changes preferentially fix or introduce errors.

3.2 BLIND DIRECTIONAL JUDGING AND THE ESTIMATOR

For each $x \in \mathcal{D}$, a judge receives the task and both candidate responses, without being told which came from the old or new model and without access to the answer label. Let $s(x) \in \{-1, 0, +1\}$ be the directional vote after mapping the randomized candidate order back to model identity: $+1$ if the judge prefers the new response, -1 if it prefers the old, and 0 for a tie. The judge's net directional preference is $J = \frac{1}{m} \sum_{x \in \mathcal{D}} s(x) = (W_n - W_o)/m$.

J carries directional information but is not on the scale of an accuracy change: a strong preference on two disagreements should not imply the same expected reliability change as the same preference on half of all probes. We therefore define the Directional Disagreement Estimator as the product $\widehat{\Delta p}_{\mathrm{DDE}} = d_{\mathrm{parsed}}\, J$, which simplifies to

$$\boxed{\widehat{\Delta p}_{\mathrm{DDE}} = \frac{W_n - W_o}{N}} \tag{6}$$

with ties contributing zero. The two factors are intended to play distinct roles: J supplies orientation and d_{parsed} maps that orientation into per-probe units. The first role is borne out empirically, the second only partially, so we present DDE as a directional statistic rather than a calibrated estimate of the magnitude of Δp (Section 5.4).

Computing DDE requires no correctness labels for the probe set. Establishing that a judge is informative on a new target domain is a separate question: Section 3.3 shows the estimator is trustworthy only when $q > 0.5$, and verifying that empirically requires either prior external validation of the judge or a labeled calibration set (Section 7).

3.3 EFFECT OF JUDGE ERROR

Consider an idealized model of judge behavior on *discriminable* disagreements, those where exactly one of the two responses is correct. Suppose the judge selects the correct response with probability q, the opposite with probability $1 - q$, and behaves symmetrically on old-correct and new-correct cases; ignore ties for the moment. For an item counted in c the expected directional vote is $q - (1 - q) = 2q - 1$, and for an item counted in b it is $-(2q - 1)$. If judge preferences on the e both-wrong disagreements have zero mean, then summing over $\mathcal{D}$ and dividing by N gives

$$\boxed{\mathbb{E}\left[\widehat{\Delta p}_{\mathrm{DDE}}\right] = (2q - 1)\,\Delta p} \tag{7}$$

under these assumptions. This yields three regimes: for $q > 0.5$ the direction is preserved in expectation; at $q = 0.5$ the expectation vanishes; and for $q < 0.5$ the direction is inverted in expectation.

The relationship is not claimed to hold exactly for real judges. Ties, position bias, asymmetric errors, stylistic preferences, and systematic preferences among both-wrong responses all introduce additional terms, and pairwise LLM judges are known to exhibit position and preference biases that make these deviations practically relevant (Shi et al., 2025). Appendix B states the assumptions in full and works through the neglected terms.

Our experiments stress-test this condition directly, with one judge measured above chance and one at chance (Section 5.5). We therefore treat above-chance judge discrimination as a precondition to check, not as an automatic property of an LLM judge.

4 EXPERIMENTAL DESIGN

Benchmark and splits. We evaluate DDE on GPQA Main, an expert-written multiple-choice benchmark of difficult graduate-level questions in physics, biology, and chemistry (Rein et al., 2024). Of its 448 questions, a manifest fixed before any model evaluation assigns 64 to an unlabeled probe pool and the 384 disjoint remainder to held-out evaluation; no item appears in both. The probe questions are the only questions available to the estimators that use no labels, and their correctness labels are hidden from DDE, the blind judge, and every other unlabeled signal. Held-out labels are used only after inference, to compute the realized Δp per transition. Because DDE contains

no learned parameters, no development split is reserved: the estimator, the probe partition, and the analysis targets were all fixed before the confirmatory evaluation. DDE was specified *after* an earlier exploratory study on MMLU-Pro and before the GPQA evaluation, so we treat GPQA as the primary evidence.

Transitions. Fourteen model versions answered all 448 questions, producing 6,272 task-model responses. From these we construct 22 primary transitions, each a distinct unordered pair, of which 16 are model replacements; the remainder are prompt and scaffold changes, retained in the confirmatory analysis rather than dropped according to their observed effect. The 16 model-replacement transitions form the subset used for judge robustness, because both judges evaluated them. Three additional reverse-direction transitions are diagnostic only. For a pair M_a, M_b the reversal tests the expected transformation properties, $\Delta p_{b\to a} = -\Delta p_{a\to b}$ for the target and $D(M_a, M_b) = D(M_b, M_a)$ for a symmetric statistic; counting both orientations as independent would inflate the effective sample size, so reverse pairs are excluded from the primary n. Appendix A lists the full version inventory and transition table.

Held-out target. Both versions are evaluated on the same 384 held-out questions, and the target is the paired accuracy difference $\widehat{\Delta p} = \widehat{p}_n - \widehat{p}_o$. Because the two versions answer identical questions, uncertainty is computed from paired outcomes rather than by treating the two accuracies as independent samples; we use McNemar-style paired intervals.

Unlabeled signals. We compare three classes. The *agreement-distance baseline* $d_{\text{agreement}}$ counts an item as agreement only when both responses parse *and* their labels match:

$$d_{\text{agreement}} = 1 - \tfrac{1}{N} \sum\nolimits_i \mathbf{1}[a_o(x_i) \neq \varnothing \wedge a_o(x_i) = a_n(x_i)]\,, \tag{8}$$

so two unparseable responses count as a disagreement. Larger values denote greater behavioral distance, and the statistic is symmetric in the two versions. *Jensen–Shannon divergence*, computed where response-distribution information is available, is likewise symmetric and so cannot by itself identify orientation. *Blind judge preference* J is obtained on the disagreement set as described in Section 3, with candidate order randomized and then mapped back to model identity; the judge sees neither the GPQA label nor an old/new identifier. DDE combines the gating rate d_{parsed} of Equation 4 with judge preference J as $d_{\text{parsed}}\, J$, with no fitted calibration parameter. The judge prompt and the randomization and parsing implementation are given in Appendix C.

The DDE gating rate d_{parsed} and the standalone baseline $d_{\text{agreement}}$ differ only when both responses are unparseable; both are symmetric under model-order reversal. We keep them distinct because the correlation attributed to the baseline in Table 1 is a property of that statistic, not of the quantity inside DDE.

Judge robustness. Two frozen judges, `gpt-5-2025-08-07` and `gpt-4o-2024-05-13`, were run on the same 16 model-replacement transitions. Judge quality is measured post hoc for mechanism analysis and is never supplied to DDE: on a disagreement where exactly one model answer is correct, which we call *discriminable*, we measure $q = P(\text{judge prefers the correct response})$. This lets us test the prediction that DDE loses directional information as q approaches 0.5. The measured values are $q = 0.657$ (Wilson 95% CI $[0.612, 0.699]$) for the first judge and $q = 0.499$ ($[0.447, 0.550]$) for the second. We additionally run an A/B order-swap pass to characterize positional consistency; it is a diagnostic only and does not alter DDE predictions. Interval and diagnostic procedures are detailed in Appendix G.

Labeled canary baselines. To quantify what unlabeled evaluation is worth relative to conventional verification, we form nested labeled canary subsets of the probe pool at $k \in \{8, 16, 32, 64\}$. For each k the labeled baseline directly estimates the accuracy difference on those k questions. The comparison is deliberately favorable to labeled evaluation: the canaries draw on the same probe budget as the unlabeled method but additionally receive ground truth. We report raw canary performance at every budget, plus a secondary labeled-question-equivalent (LQE) analysis estimating the labeled budget at which canary rank correlation matches DDE, always with bootstrap uncertainty rather than as a single exact number (Appendix E).

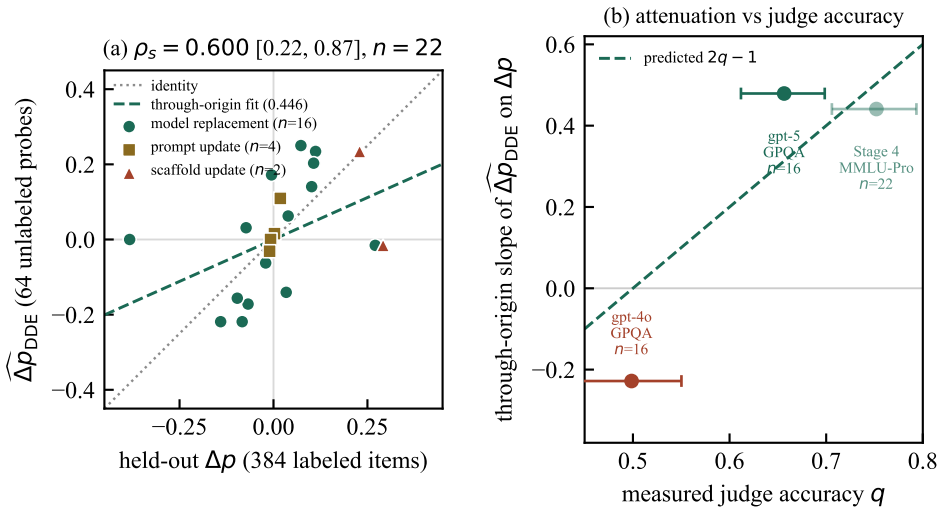


Figure 1: **(a)** DDE against held-out signed reliability change, one point per primary transition. Dashed line: through-origin fit; dotted line: identity. **(b)** Through-origin attenuation slope against measured judge accuracy q, with Wilson 95% intervals on q. All slopes use the same estimator; three operating points cannot identify the functional form. The Stage-4 MMLU-Pro point is post-hoc exploratory.

Metrics. We report the Spearman rank correlation $\rho_s(S, \Delta p)$ between each signal S and held-out accuracy change, and $\rho_s(S, |\Delta p|)$ to separate direction from magnitude. For DDE we add sign accuracy and mean absolute error, each against a trivial baseline, and evaluate detection of $\Delta p < -0.05$ by AUROC. All primary analyses treat the transition, not the benchmark question, as the statistical unit.

5 Results

5.1 DDE predicts signed reliability change

The pre-specified confirmatory GPQA evaluation supports the primary hypothesis. Across the 22 primary transitions, DDE correlates with held-out signed accuracy change at $\rho_s = 0.600$ (95% CI $[0.22, 0.87]$, $p = 0.0031$); a secondary sign-accuracy comparison gives 0.68 against a 0.50 constant-sign baseline (Section 5.4, Appendix F). For detecting updates that reduce accuracy by more than five percentage points — 6 of the 22 transitions, a prevalence of 0.273 — DDE attains AUROC 0.849 (95% CI $[0.611, 1.000]$); see Figure 1.

With only 22 primary transitions uncertainty remains substantial, and the AUROC interval reaches close to chance at its lower bound, so we treat the screening figure as indicative rather than as a precise operating characteristic. The association is not driven by a single model endpoint: leave-one-endpoint-out correlations remain positive across all 14 endpoints (0.496–0.753, median 0.599); an endpoint-aware vertex-bootstrap sensitivity interval is $[0.146, 0.899]$ (Appendix F).

5.2 Direction and magnitude separate empirically

Table 1 summarizes the central empirical dissociation. The symmetric agreement-distance baseline $d_{\mathrm{agreement}}$ is essentially unrelated to signed accuracy change but substantially associated with the absolute change, and the judge-based measurements show exactly the complementary pattern.

These are therefore not weaker and stronger versions of a single notion of model change: symmetric difference carries extent, comparative preference carries orientation. JS divergence is less clean, showing substantial association with magnitude (0.743) alongside a moderate signed correlation (0.434); being symmetric by construction, however, any signed association in a finite set of one-way transitions cannot provide orientation consistently under reversal.

Table 1: Direction and magnitude are carried by different signals. Rank correlations against signed held-out reliability change and against its absolute value, over the 22 primary GPQA transitions.

Signal	Probe labels	ρ vs. Δp	ρ vs. $\lvert\Delta p\rvert$
DDE	0	**0.600**	0.028
Judge preference J	0	**0.588**	0.021
Agreement distance $d_{\text{agreement}}$	0	-0.060	**0.565**
JS divergence	0	0.434	**0.743**
64-item canary	64	**0.850**	0.134

5.3 Reverse transitions verify the symmetry constraint

Reverse-pair diagnostics reproduce the structural result in the evaluation pipeline: both symmetric disagreement definitions are unchanged under reversal, while Δp changes sign. Full pairwise values are reported in Appendix D.

5.4 Directional utility survives trivial baselines; absolute calibration does not

Checked against trivial predictors, the two halves of the estimator come apart.

The directional claim survives. The 22 primary transitions are balanced, 11 improvements and 11 regressions, so every constant-sign rule — always improvement, always regression, or the majority sign — scores exactly 0.500, against 0.682 for DDE.

The calibration claim does not. Multiplying J by the gating rate d_{parsed} reduces absolute error relative to raw J (MAE 0.1115 versus 0.2306), but does not beat the trivial zero-change predictor $\widehat{\Delta p} = 0$, which attains MAE 0.0990 and RMSE 0.1425 against DDE's 0.1115 and 0.1512. The mechanism is visible in two quantities already reported: the regression slope is below one (0.4463) while the ratio of mean magnitudes exceeds one (1.148), so DDE overshoots the size of a change while under-responding to it. DDE also only slightly improves rank correlation over raw judge preference (0.600 against 0.588).

We therefore claim neither that $d_{\text{parsed}}J$ is a better *ranking* statistic than the judge alone, nor that it is a calibrated estimate of magnitude. DDE identifies update direction more accurately than the constant-sign baselines on this balanced set, but does not yet say how many accuracy points it moved better than assuming no change.

5.5 The judge-quality condition separates success from failure

The two-judge comparison directly tests the dependence predicted in Section 3.3. On the same 16 model-replacement transitions, the GPT-5 judge has discriminative accuracy $q = 0.657$, and its DDE estimates correlate with the true reliability changes at $\rho_s = 0.651$ ($p = 0.0063$), with sign accuracy 0.688 and a through-origin DDE–Δp slope of 0.4791. The GPT-4o judge is measured almost exactly at chance, $q = 0.499$; with this judge the DDE correlation falls to $\rho_s = -0.140$ ($p = 0.604$) and sign accuracy to 0.312.

The second judge therefore does not establish judge-independence. It shows the opposite: useful DDE estimates depend on the comparative judge carrying genuine task-relevant information. This failure is consistent with the simplified theoretical model, in which $2q - 1$ approaches zero as $q \approx 0.5$ and the expected directional signal is attenuated away. A post-hoc mechanism decomposition supports this interpretation: the capable judge reverses preference according to which model is correct, whereas the chance-level judge exhibits nearly the same directional tendency in both correctness orientations; detailed counts are in Appendix B.

The 16 model-replacement transitions, taken alone, give $\rho_s = 0.651$ (95% CI $[0.218, 0.900]$). The remaining subgroups are too small to analyse: 4 prompt updates and 2 scaffold updates. We therefore cannot establish whether DDE generalizes equally across update types, and report no significance tests for those subgroups (Appendix F).

Across GPQA and the earlier exploratory MMLU-Pro analysis, the observed through-origin attenuation slopes are 0.4463 and 0.4409 respectively, when computed with the same estimator. Figure 1(b) plots the three operating points. We treat them as consistent with, rather than establishing, the proposed dependence on judge accuracy; the number of judge/benchmark conditions is far too small to identify a general functional relationship.

5.6 Judge consistency is not judge correctness

Order consistency was similar for the informative and chance-level judges, showing that consistency does not imply correctness; DDE also exhibited imperfect antisymmetry under reversal (Appendix D). Position consistency is therefore not a substitute for measuring judge discrimination on a small labeled calibration set.

5.7 Labels outperform the unlabeled estimator

Labeled evaluation remains stronger: 32 labeled canaries reach $\rho_s = 0.847$, while DDE's estimated labeled-question equivalent is approximately 16, with a wide bootstrap CI of $[8.0, 31.3]$ (Appendix E). DDE is therefore most relevant when labels are unavailable, costly, or when an inexpensive screen precedes full reevaluation.

Summary. Behavioral change alone does not determine whether an update is beneficial. Across the confirmatory GPQA transitions, symmetric response differences primarily track the amount of change, while a capable comparative judge provides directional information. Combining the two yields a probe-label-free directional score, but the estimator succeeds only when the judge itself is informative, it does not beat a zero-change predictor on absolute error, and it remains weaker than modest labeled evaluation.

6 Related Work

Model-update monitoring and degradation testing. Recent work studies whether an LLM endpoint has changed and whether a measured performance drop is statistically meaningful. Leshin et al. (2026) introduce behavioral fingerprints for black-box endpoint stability, comparing output distributions from fixed prompts to detect changes in model family, version, inference stack, quantization, and sampling behavior. Their objective is change detection rather than determining whether the change improves or degrades task reliability. Kübler et al. (2026) instead address degradation directly, proposing paired statistical tests based on per-example correctness and McNemar's test. Their framework assumes labeled evaluation outcomes and asks whether an observed accuracy drop reflects genuine degradation rather than sampling noise. Our setting lies between these: we observe paired model responses but ask what can be inferred about the *sign* of reliability change before using ground-truth labels.

Estimating performance from unlabeled data. A long line of work estimates model error without labels. Donmez et al. (2010) estimate classification and regression error from unlabeled data, Jaffe et al. (2015) recover the accuracies of multiple classifiers without labeled data, and Platanios et al. (2016) give a Bayesian treatment of the same problem. These methods target the *absolute* accuracy or error of classifiers, typically by exploiting agreement structure across an ensemble of predictors under independence or other modeling assumptions. Our target is different in both the estimand and the mechanism: we estimate the *sign* of a paired old-to-new reliability change for a single evolving system, and we obtain orientation from comparative judgments of the two responses rather than from agreement structure across many predictors. Lee et al. (2023) analyze disagreement-on-the-line, showing that disagreement between models carries real information about error under distribution shift. Our symmetry result does not contradict this and does not claim that disagreement is uninformative; it says only that a *symmetric* disagreement statistic cannot by itself encode orientation under reversal of the pair, which is consistent with disagreement being highly informative about magnitude, as we in fact observe.

LLM-as-a-judge and regression evaluation. LLM judges are widely used to compare model outputs and have also been incorporated into regression-testing systems. Liusie et al. (2024) use

pairwise LLM judgments for zero-shot NLG evaluation and analyze positional effects, and GRAFITE evaluates model issues across releases with LLM-as-a-judge to support continuous regression analysis (Lee et al., 2026). We do not claim comparative judging as a new primitive, nor do we propose a new way to rank models against one another. Instead, we study the judge's role as an *orientation-bearing signal* attached to a symmetric magnitude signal, for the specific target of a paired old-to-new reliability change. This distinction matters because LLM judges themselves exhibit systematic biases. Shi et al. (2025) document position bias and varying consistency across judges and tasks in large-scale pairwise evaluation, and Jeong et al. (2025) show that pairwise comparison can amplify evaluators' preferences for superficial features rather than correctness. Our order-swap diagnostics, imperfect antisymmetry, and chance-level second judge all reinforce this limitation: DDE is useful only when the judge carries task-relevant discriminative information.

7 LIMITATIONS

Our conclusions are intentionally narrower than a general solution to model regression detection. First, the 22 primary transitions reuse 14 underlying model versions and are dominated by models from one provider; transition-level analyses therefore provide more variation than the number of underlying model families. Second, GPQA is the only pre-specified confirmatory benchmark. The earlier MMLU-Pro study is exploratory and should not be treated as an independent confirmatory replication.

Third, DDE relies on the comparative judge. The second judge performs at chance on discriminable response pairs and yields no useful directional signal, showing that judge capability must be checked rather than assumed. Pairwise judging also exhibits order and style sensitivity — including preferences for superficial features such as verbosity and confident phrasing (Jeong et al., 2025) — and DDE is not perfectly antisymmetric under transition reversal.

This qualification bounds the method's practical value. DDE itself requires no labels for the update probe set, but establishing that the judge is above chance does require labels somewhere. If judge suitability must be re-established with fresh in-domain labels for every update, direct labeled canary evaluation may be the more efficient use of that labeling budget. Our experiments do not establish how reliably a judge's qualification transfers across model updates or across domains, and we regard this as the most important open question for deploying the method.

Fourth, the evidence supports DDE as a directional statistic only: it does not outperform a zero-change predictor on absolute error (Section 5.4), so its magnitude should not be read as a calibrated estimate. The 16 model-replacement transitions also dominate the primary set; 4 prompt and 2 scaffold updates are too few to establish whether the estimator behaves equally across update types.

Fifth, our experiments use objectively scored multiple-choice tasks, where disagreement is easy to define and held-out correctness provides a clear Δp; open-ended generation, tool use, or subjective quality criteria may require different notions of disagreement and reliability. Finally, labeled evaluation remains stronger when trustworthy labels are available: DDE is a screening signal for settings where labels are costly or unavailable, not a replacement for labeled regression testing.

8 CONCLUSION

Behavioral change does not by itself reveal whether an LLM update is beneficial. Symmetric change statistics can characterize the extent of a shift but cannot consistently encode its orientation. We introduced DDE, which combines response disagreement with a blind comparative judge to estimate signed reliability change from unlabeled probes. On the pre-specified GPQA evaluation, DDE provides useful directional signal when the judge is demonstrably above chance and collapses when the judge is not, though it does not outperform a zero-change predictor on absolute-error metrics. These results suggest a practical separation between *detecting how much a model changed* and *determining which way reliability moved*, while also making clear that modest labeled evaluation remains preferable whenever reliable labels are available.

AI USE STATEMENT

Generative AI tools were used to assist with literature exploration, research-question refinement, experimental design, software implementation and debugging, statistical-analysis workflows, mathematical reasoning, interpretation of results, and drafting and editing portions of the manuscript. All experimental results reported in this paper were produced by the reported code and model endpoints, not generated by an assistant. The author reviewed and verified the analyses, references, numerical claims, mathematical arguments, and final manuscript, and takes responsibility for the work.

REFERENCES

Pinar Donmez, Guy Lebanon, and Krishnakumar Balasubramanian. Unsupervised supervised learning i: Estimating classification and regression errors without labels. *Journal of Machine Learning Research*, 11(44):1323–1351, 2010. URL https://www.jmlr.org/papers/v11/donmez10a.html.

Ariel Jaffe, Boaz Nadler, and Yuval Kluger. Estimating the accuracies of multiple classifiers without labeled data. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *Proceedings of Machine Learning Research*, pp. 407–415, 2015. URL https://proceedings.mlr.press/v38/jaffe15.html.

Hawon Jeong, ChaeHun Park, Jimin Hong, Hojoon Lee, and Jaegul Choo. The comparative trap: Pairwise comparisons amplifies biased preferences of LLM evaluators. In *Proceedings of the 8th BlackboxNLP Workshop*, pp. 79–108, 2025. URL https://aclanthology.org/2025.blackboxnlp-1.5/.

Jonas Kübler, Kailash Budhathoki, Matthäus Kleindessner, Xiong Zhou, Junming Yin, Ashish Khetan, and George Karypis. When LLMs get significantly worse: A statistical approach to detect model degradations. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/70de9e3948645a1be2de657f14d85c6d-Abstract-Conference.html.

Donghwan Lee, Behrad Moniri, Xinmeng Huang, Edgar Dobriban, and Hamed Hassani. Demystifying disagreement-on-the-line in high dimensions. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19053–19093, 2023. URL https://proceedings.mlr.press/v202/lee23o.html.

Ja Young Lee, Mírian Silva, Mohamed Nasr, Shonda Witherspoon, Enzo Bozzani, Veronique Demers, Radha Ratnaparkhi, Hui Wu, and Sara Rosenthal. GRAFITE: Generative regression analysis framework for issue tracking and evaluation. *arXiv preprint*, 2026. URL https://research.ibm.com/publications/grafite-generative-regression-analysis-framework-for-issue-tracking-and-evaluation.

Jonah Leshin, Manish Shah, Ian Timmis, and Daniel Kang. Behavioral fingerprints for LLM endpoint stability and identity. In *Proceedings of the ACM Conference on AI and Agentic Systems (CAIS)*. Association for Computing Machinery, 2026. doi: 10.1145/3786335.3813194.

Adian Liusie, Potsawee Manakul, and Mark Gales. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 139–151, 2024. URL https://aclanthology.org/2024.eacl-long.8/.

Emmanouil Antonios Platanios, Avinava Dubey, and Tom Mitchell. Estimating accuracy from unlabeled data: A Bayesian approach. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1416–1425, 2016. URL https://proceedings.mlr.press/v48/platanios16.html.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2311.12022.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the Joint Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (IJCNLP-AACL)*, 2025. URL https://aclanthology.org/2025.ijcnlp-long.18/.

A Full transition inventory

Table 2 lists the 14 evaluated versions. Every endpoint is an exact dated snapshot rather than a floating alias, and the provider's reported model id was recorded on each response and checked against the requested id, so a silent re-route would have been visible in the data. Versions differ along three axes: the underlying snapshot, the system prompt, and the decoding or reasoning scaffold. Where a version specifies no explicit output budget the pipeline default of 512 tokens applies.

Table 3 gives the 22 primary transitions with their disagreement counts, judge preferences, DDE values and held-out targets, and Table 4 the 3 reverse-direction diagnostics. Primary transitions are distinct unordered version pairs; the diagnostics are the reversals of `M01`, `M07` and `P01` and are excluded from every cross-transition statistic so that no pair contributes twice to the effective sample size.

Table 2: The 14 evaluated model versions. All endpoints are exact dated snapshots. `effort` is the `reasoning_effort` parameter where the endpoint accepts one. Temperature is 0 for every version. Prompt and scaffold labels are shortened; full identifiers are in the configuration.

Version	Snapshot	Prompt	Scaffold	Max tok.	Effort
v_full41	gpt-4.1-2025-04-14	neutral	—	512	—
v_full4o	gpt-4o-2024-08-06	neutral	—	512	—
v_mini41	gpt-4.1-mini-2025-04-14	neutral	—	512	—
v_mini41_concise	gpt-4.1-mini-2025-04-14	concise	—	512	—
v_mini4o	gpt-4o-mini-2024-07-18	neutral	—	512	—
v_mini4o_concise	gpt-4o-mini-2024-07-18	concise	—	512	—
v_mini4o_speed	gpt-4o-mini-2024-07-18	speed	—	512	—
v_mini4o_structured	gpt-4o-mini-2024-07-18	structured	—	512	—
v_mini5	gpt-5-mini-2025-08-07	neutral	minimal	512	minimal
v_mini54	gpt-5.4-mini-2026-03-17	neutral	low	2048	low
v_mini5_low	gpt-5-mini-2025-08-07	neutral	low	2048	low
v_nano41	gpt-4.1-nano-2025-04-14	neutral	—	512	—
v_nano5	gpt-5-nano-2025-08-07	neutral	minimal	512	minimal
v_nano5_low	gpt-5-nano-2025-08-07	neutral	low	8192	low

B Extended derivation and judge-mechanism decomposition

Scope of Proposition 1. The proposition rules out one specific thing: recovering the *sign* of Δp from a symmetric statistic alone, for both orientations of the same pair. It does not say that symmetric statistics are uninformative, and our own results show the opposite for magnitude. It also does not apply to a statistic that is given side information breaking the symmetry — a timestamp, a version ordering, or a directional judgment such as J. The estimator in this paper is precisely such a construction: a symmetric magnitude term multiplied by an antisymmetric orientation term.

Assumptions behind Equation 7. The derivation assumes (i) the judge picks the correct response with a single probability q on discriminable disagreements, (ii) that probability is the same whether the old or the new version is the correct one, (iii) judge preferences on both-wrong disagreements have mean zero, and (iv) ties are absent. Each is an idealisation. Assumption (ii) fails whenever the judge has a standing preference for one side, which is exactly the failure mode we observe in the chance-level judge below. Assumption (iv) is mild here: tie rates are 0.4% (`gpt-5`) and 1.4% (`gpt-4o`), and ties enter J's denominator while contributing nothing to its numerator, which shrinks the estimate slightly toward zero. Position bias, verbosity preference and stylistic preference all add

Table 3: The 22 primary transitions. m is the number of probe disagreements under the parsed-answer definition, $d_{\text{parsed}} = m/64$, J the judge's net preference and Δp the held-out target on 384 labeled items. Every disagreement in this table was judged, so $n_{\text{judged}} = m$ for all 22 rows.

ID	Old	New	Update type	m	d	J	DDE	Δp
M01	v_nano41	v_mini41	model	29	0.453	0.552	0.250	0.073
M02	v_nano41	v_full41	model	37	0.578	0.405	0.234	0.112
M03	v_nano5	v_mini5	model	35	0.547	0.371	0.203	0.107
M04	v_mini4o	v_mini41	model	41	0.641	0.220	0.141	0.102
M05	v_mini41	v_full41	model	24	0.375	0.167	0.062	0.039
M06	v_mini5	v_mini54	model	34	0.531	-0.029	-0.016	0.271
M07	v_full41	v_mini4o	model	36	0.562	-0.389	-0.219	-0.141
M08	v_full4o	v_mini4o	model	31	0.484	-0.355	-0.172	-0.068
M09	v_mini41	v_nano5	model	41	0.641	-0.244	-0.156	-0.096
M10	v_mini5	v_nano41	model	37	0.578	-0.378	-0.219	-0.083
M11	v_mini54	v_mini4o	model	42	0.656	0.000	0.000	-0.383
M12	v_full4o	v_mini41	model	28	0.438	-0.321	-0.141	0.034
M13	v_nano5	v_mini4o	model	43	0.672	0.256	0.172	-0.005
M14	v_mini5	v_mini41	model	27	0.422	-0.037	-0.016	-0.010
M15	v_full41	v_full4o	model	21	0.328	0.095	0.031	-0.073
M16	v_mini5_low	v_mini54	model	18	0.281	-0.222	-0.062	-0.021
P01	v_mini4o	v_mini4o_concise	prompt	17	0.266	0.412	0.109	0.018
P02	v_mini4o	v_mini4o_speed	prompt	11	0.172	0.091	0.016	0.003
P03	v_mini4o	v_mini4o_structured	prompt	16	0.250	0.000	0.000	-0.008
P04	v_mini41	v_mini41_concise	prompt	18	0.281	-0.111	-0.031	-0.010
S01	v_mini5	v_mini5_low	scaffold	28	0.438	-0.036	-0.016	0.292
S02	v_nano5	v_nano5_low	scaffold	35	0.547	0.429	0.234	0.229

Table 4: The 3 reverse-direction diagnostic transitions, excluded from every cross-transition statistic and from the primary n. m is the parsed-answer disagreement count and n_{judged} how many of those a judge scored. Judging of two diagnostics was cut short when API credit was exhausted, so $n_{\text{judged}} < m$ there. Those rows use a diagnostic partial-coverage extension: J is estimated on the judged subset and multiplied by the full d_{parsed}, extrapolating the observed preference rate to the full disagreement set. All 22 primary transitions have complete judge coverage and use the estimator exactly as defined in Section 3.2.

ID	Old	New	m	d	n_{judged}	J	DDE	Δp
D01_REV_M01	v_mini41	v_nano41	29	0.453	29	-0.483	-0.219	-0.073
D02_REV_M07	v_mini4o	v_full41	36	0.562	31	0.097	0.054	0.141
D03_REV_P01	v_mini4o_concise	v_mini4o	17	0.266	12	0.167	0.044	-0.018

terms that do not cancel under reversal, and are the most plausible explanation for the imperfect antisymmetry reported in Appendix D.

Judge behaviour by hidden-label group. Correctness labels were read post hoc for this analysis and were never supplied to the judge or to DDE. Table 7 decomposes every primary-pass verdict by which version was actually correct. The capable judge reverses its mean directional vote with correctness, while the chance-level judge votes in the same direction in both orientations — a constant mild preference for the newer response that carries no information about which answer is right. This is the mechanism behind the ρ_s collapse reported in Section 5.5.

The both-wrong zero-mean assumption. Equation 7 assumes judge preferences on both-wrong disagreements average to zero. Those items are 35.6% (`gpt-5`) and 36.3% (`gpt-4o`) of judged disagreements, so a systematic preference among them would matter. The assumption is supported in this experiment: the mean directional vote on both-wrong items is -0.0156, 95% transition-clustered bootstrap CI $[-0.197, 0.159]$, for `gpt-5`, and -0.0048, $[-0.207, 0.187]$, for `gpt-4o`. Both intervals contain zero. Source: `results/paper/reviewer_audit.json`, section E.

Table 5: Leave-one-endpoint-out sensitivity. Each row removes every transition touching that endpoint and recomputes the DDE–Δp rank correlation on the remainder.

Endpoint removed	Transitions left	ρ_s	p
`v_mini5`	17	0.753	0.000
`v_mini54`	19	0.727	0.000
`v_mini5_low`	20	0.678	0.001
`v_full4o`	19	0.622	0.004
`v_mini4o`	14	0.615	0.019
`v_mini4o_speed`	21	0.603	0.004
`v_mini4o_structured`	21	0.602	0.004
`v_mini4o_concise`	21	0.596	0.004
`v_nano5_low`	21	0.580	0.006
`v_mini41_concise`	21	0.577	0.006
`v_full41`	18	0.561	0.015
`v_nano5`	18	0.516	0.028
`v_nano41`	19	0.510	0.026
`v_mini41`	15	0.496	0.060

Table 6: Reverse-pair diagnostics. Both disagreement definitions are exactly unchanged under reversal — $d_{\text{agreement}}$, which counts two unparseable answers as disagreeing, and d_{parsed}, which counts them as agreeing — while Δp reverses exactly. DDE does neither; the last column is the antisymmetry residual.

	$d_{\text{agreement}}$		d_{parsed}		Δp		DDE		
Pair	fwd	rev	fwd	rev	fwd	rev	fwd	rev	|sum|
`M01 / D01_REV_M01`	0.484	0.484	0.453	0.453	0.073	-0.073	0.250	-0.219	0.031
`M07 / D02_REV_M07`	0.562	0.562	0.562	0.562	-0.141	0.141	-0.219	0.054	0.164
`P01 / D03_REV_P01`	0.266	0.266	0.266	0.266	0.018	-0.018	0.109	0.044	0.154

C JUDGE PROMPT AND IMPLEMENTATION

The judge is given the question, the option list, and the two candidate responses, with no indication of which system produced either and no access to the answer key. The system prompt is:

```
You are comparing two candidate answers to a multiple-choice
question.  Decide which candidate is more likely to be
correct.  You are not told which system produced either
candidate.  Reply with exactly one token:  A, B, or TIE.
```

The user message presents `Question`, `Options`, `Candidate A` and `Candidate B` as labelled blocks.

Order randomization. Whether the new version's response appears as Candidate A is decided per item by a seeded permutation, recorded as `new_was_a`, and inverted when the verdict is decoded back to a directional vote. A separate order-swap pass re-runs a subset with the candidates exchanged; those verdicts are recorded with `swapped_pass=true` and are used only for the position-bias diagnostic, never in DDE, so that no item is double-counted.

Parsing and degenerate verdicts. The verdict is parsed as `A`, `B` or `TIE`; anything else is recorded as `UNPARSEABLE`. Ties and unparseable verdicts contribute zero to the numerator of J but remain in its denominator. Judging is run only on the disagreement set, and the judge is queried at temperature 0 with a fixed seed. Because the cache is content-addressed, two transitions that present the judge with an identical pair of response texts share one paid call; the 1,611 recorded verdicts therefore come from 1,017 distinct judge queries. Verdicts are unique per (judge, transition, item, pass).

Table 7: Judge behaviour by hidden-label group on GPQA probe disagreements. Mean vote is $+1$ for preferring the new response, -1 the old, 0 a tie. Accuracy is computed only where exactly one version is correct.

Judge	Group	n	Mean vote	Prefers new	Accuracy
`gpt-5`	new correct, old wrong	251	$+0.327$	0.653	0.667
	old correct, new wrong	213	-0.286	0.352	0.645
	both wrong	257	-0.016	0.490	—
`gpt-4o`	new correct, old wrong	178	$+0.062$	0.511	0.532
	old correct, new wrong	190	$+0.063$	0.521	0.468
	both wrong	210	-0.005	0.486	—

D REVERSE-PAIR DIAGNOSTICS AND JUDGE ORDER CONSISTENCY

Two disagreement definitions. Both symmetric definitions appear in Table 6, because they differ on one edge case. $d_{\text{agreement}}$, the standalone baseline of Equation 8, counts an item as agreeing only when both answers parse *and* match, so two unparseable responses count as a disagreement; d_{parsed}, the gating rate inside DDE (Equation 4), counts them as agreeing. They coincide except on `M01`, where two items have both responses unparseable (0.484 versus 0.453). Neither definition depends on the order of the two versions, and both are exactly unchanged under reversal.

Symmetry and antisymmetry. Over the three reverse pairs the symmetric statistics are unchanged to machine precision, $\max |d_{\text{agreement}}(M_a, M_b) - d_{\text{agreement}}(M_b, M_a)| = 0$ (the same holds for d_{parsed} and for JS divergence), while Δp reverses sign exactly. DDE is only approximately antisymmetric: it flips sign on two of the three pairs, with mean $|\text{DDE}_{a\to b} + \text{DDE}_{b\to a}| = 0.116$. This is a property of the judge-based estimator, not of the symmetry argument of Proposition 1.

Judge coverage on the diagnostics. Judging of the reverse diagnostics was interrupted when API credit was exhausted, and the salvage pass recorded only the comparisons already paid for. Two diagnostics are therefore partially judged: `D02_REV_M07` has 31 verdicts for 36 disagreements and `D03_REV_P01` has 12 for 17. Because these reverse diagnostics have incomplete judge coverage, we use a diagnostic partial-coverage extension: J is estimated on the judged subset and multiplied by the full parsed-answer disagreement rate d_{parsed}, so the reported DDE remains on the per-probe scale. This extrapolates the observed preference rate to the full disagreement set and is used only for the reverse-pair diagnostic; all 22 primary transitions have complete judge coverage and use the estimator exactly as defined in Section 3.2. Using the judged count n_{judged} in place of d_{parsed} instead would move the mean antisymmetry residual from 0.116 to 0.115, which does not change the conclusion. **No primary transition is affected**: all 22 have every disagreement judged. Source: `results/paper/reverse_pair_audit.json`.

Order consistency. Across the A/B order-swap pass, the fraction of swap pairs on which a judge keeps the same preference is 0.684 for `gpt-5` and 0.664 for `gpt-4o`. These are close despite discriminative accuracies of 0.657 and 0.499 respectively, which is the basis for the claim in Section 5.6 that positional consistency does not imply correctness. Source: `results/paper/reviewer_audit.json`.

E CANARY LADDER AND LQE PROCEDURE

Construction. The canary subsets are nested prefixes of the same 64 probe questions, at $k \in \{8, 16, 32, 64\}$, so a larger budget always contains the smaller ones and the ladder is monotone in information. Crucially the canaries reuse the probe responses already collected, so they cost no additional model calls and differ from the unlabeled signals only in receiving ground truth. For each k the labeled baseline is the accuracy difference between the two versions on those k items.

Ladder. Rank correlation with held-out Δp at labeled budgets $k \in \{8, 16, 32, 64\}$ is 0.555, 0.604, 0.847 and 0.850 respectively. Thirty-two labeled questions already reach AUROC 0.98 for detecting $\Delta p < -0.05$, substantially above DDE's 0.849.

LQE. The observed ladder is first monotonised by running maximum, so a rung that happens to dip below a smaller budget does not create a spurious crossing. The unlabeled signal's quality is then located on that curve and the equivalent budget read off by interpolating in $\log_2 k$, since the ladder is approximately linear in log budget over this range. Uncertainty comes from a bootstrap that resamples *transitions* and recomputes both the signal quality and the entire ladder inside each replicate, so the interval reflects the paired noise rather than treating the two curves as independent. Interpolating DDE's quality ($\rho_s = 0.600$) gives a labeled-question equivalent of 15.16 questions, with interval $[8.0, 31.3]$. The lower bound sits at the smallest rung tested, so the interval is informative mainly on the upper side, and we treat LQE only as a secondary practitioner-facing interpretation rather than as a precise exchange rate.

F Additional robustness analyses

Endpoint dependence. The 22 primary transitions are built from 14 endpoints, so they are not independent units and a plain transition bootstrap will be optimistic. Table 5 reports leave-one-endpoint-out sensitivity: removing every transition that touches a given endpoint and recomputing the correlation on the remainder. All 14 holdouts stay positive, ranging from 0.496 to 0.753 with median 0.599, so no single endpoint drives the association. As a dependence-aware sensitivity analysis we also run a vertex bootstrap: resample the 14 endpoints with replacement, retain the transitions whose *both* endpoints were drawn, and recompute Spearman. Over 3,264 usable replicates this gives median 0.643 and a 95% interval of $[0.146, 0.899]$, wider than the transition bootstrap's $[0.217, 0.869]$ but still excluding zero. The procedure is approximate — replicate size varies and isolated vertices drop edges — so we report it as a sensitivity analysis rather than as the primary interval.

Transition-type subgroups. Model replacement: $n = 16$, $\rho_s = 0.651$, 95% CI $[0.218, 0.900]$, sign accuracy 0.688. Prompt updates: $n = 4$, sign accuracy 0.75. Scaffold updates: $n = 2$, sign accuracy 0.50. We report no correlation or significance test for the latter two: at $n = 4$ and $n = 2$ any such statistic would be uninterpretable. The consequence is that our evidence speaks to model replacement, and does not establish whether DDE behaves equally on prompt and scaffold changes.

Trivial baselines. The 22 primary transitions split 11 improvements and 11 regressions, so every constant-sign rule scores exactly 0.500 against DDE's 0.682. DDE orients 15 of the 22 correctly (Wilson 95% CI $[0.47, 0.84]$; one-sided exact binomial against 0.5, $p = 0.067$). We therefore report sign accuracy as a descriptive secondary comparison, with the rank correlation of Section 5.1 as the primary inferential result. For absolute error, the zero-change predictor attains MAE 0.0990 and RMSE 0.1425; DDE attains 0.1115 and 0.1512; bare J attains MAE 0.2306. DDE therefore beats the directional baselines and loses to the magnitude baseline, which is the basis for the framing in Section 5.4.

Screening detail. At the $\Delta p < -0.05$ threshold there are 6 positive and 16 negative transitions, a prevalence of 0.273. DDE attains AUROC 0.849 (95% CI $[0.611, 1.000]$) and AUPRC 0.808 (95% CI $[0.456, 1.000]$). The $\Delta p < -0.10$ threshold has only 2 positives; we compute it for completeness in the artifact but do not report it as an operating characteristic, because with two positive cases the estimate is not interpretable. Source: `results/paper/reviewer_audit.json`, section B.

G Evaluation and instrumentation details

Intervals. Judge accuracy q is a binomial proportion over discriminable disagreements and is reported with a Wilson score interval, which behaves better than the normal approximation near the chance boundary that matters here. The held-out Δp is a paired quantity on identical items, so its interval is McNemar-style and depends only on the discordant items; treating the two accuracies as independent samples would overstate the uncertainty. Cross-transition statistics bootstrap over transitions, with the endpoint-aware variant in Appendix F.

Dataset handling. GPQA Main is loaded once and partitioned by a manifest of item ids fixed before any model call. Answer options are shuffled with a per-item seed so that the correct answer's position is not predictable from the source ordering, and the manifest stores ids only, never question text. The probe/held-out split is disjoint by construction and verified as such in the test suite.

Caching and cost. Every model and judge response is stored in a content-addressed cache keyed by the full request, so re-runs are free and the experiment is replayable offline. Total incremental spend for the Stage-5 experiment was $3.93, reconstructed from cached token counts and the price table in the configuration.

Truncated reasoning responses. Reasoning models that exhaust their output budget on hidden reasoning return a successful response with an empty completion, which a naive parser scores as an incorrect answer. Because non-reasoning models emit the answer first, this penalty falls only on reasoning versions and is a per-version bias rather than shared noise. We detect the condition and escalate that single request's budget until the model produces content, recording the granted budget on the response; the configured budget is left unchanged, so cached responses remain valid. No such truncated response remains in the Stage-5 data.

H REPRODUCIBILITY

Every number reported in this paper is persisted in `results/paper/paper_metrics.json`, each entry carrying its producing script, source files, dataset, split and SHA-256 content hashes. A claim ledger at `results/provenance/claim_ledger.json` audits those entries and marks any whose sources have drifted. The artifact set the paper is written against is frozen in `results/paper/paper_freeze_manifest.json`, which records a SHA-256 for every input, figure and table; a verification mode re-hashes them and reports any change. The main-paper figure and every appendix table in this document are generated from those artifacts by `scripts/make_paper_figures.py` and `scripts/make_appendix_tables.py`, so they cannot silently diverge from the recorded values.