
The Screen Decides the Verdict: Operating Characteristics of Model-Exclusion Rules in LLM Evaluation

Emile Boullineau
Independent researcher
emile.boullineau@gmail.com

José Daniel Muñoz Arciniegas
Independent researcher
josed.munoz.a@gmail.com

Abstract

An evaluation of hosted language models can screen models for run-to-run stability and then vote among the survivors. We audit a preregistered evaluation in which this two-step pipeline determined the result. Four models answered a yes/no concern question about 20 behaviours written to be ambiguous, each attributed to a neutral actor or to one of six personas. The registered stability gate required a Pearson correlation above .80 between two runs over 14 conditions (two scenarios under seven frames) with 20 calls each. It excluded the model whose rates moved least on the rate scale (mean absolute change .018) and retained the model whose rates moved most (.129). A simulated model with the excluded model’s rates and no drift failed the gate in 83.9% of bootstrap draws. Across 252 simulated designs, the gate excluded models that had not drifted at a median rate of .763. At the study’s design, with an assumed spread between conditions of .5 logits, we recalibrated the threshold to a 5% false-exclusion rate. Power against an upward one-logit shift at a baseline rate of .50 (moving it to about .73) was .066. A registered hypothesis, that persona frames make a model’s answers less consistent within a condition, was to be supported if at least 3 of the 4 models showed this. After the exclusion this rule required 3 of 3, and the realised outcome, two passes and one failure, met neither the support rule nor the paired rule that would refute it. A non-confirmatory sensitivity analysis that restored the excluded model met the support rule. We study one protocol and make no prevalence claim. We show that fixed-count support and refutation rules that give every full-panel outcome exactly one verdict leave some outcomes with no verdict after a removal, and that restating the registered thresholds as fractions of survivors does not repair our case. Screens and vote rules should therefore be specified and simulated together.

1 Introduction

Some evaluations of hosted language models decide which systems to analyse before analysing them. A model can be dropped for failing a format check or refusing, or, as in the preregistered evaluation we audit here (Section 2), for disagreeing with itself across runs, and a verdict is then computed over the models that remain. The field reports variance across seeds, prompts, and formats, benchmark sensitivity, and uncertainty [1, 4, 5, 7, 14, 15, 19–21, 23]. A screen and a vote, however, form one decision procedure, and validating each stage separately does not validate the pair. We call any rule that removes models before the analysis a screen, and the registered correlation screen of our study the gate. By drift we mean a change in a model’s underlying Yes rates between two runs, as distinct from the chance variation among its stochastic calls.

Preprint note. Accepted as a poster at the TAE workshop at NeurIPS 2026 (non-archival). Camera-ready version, revised after workshop review. It replaces the prior version.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: TAE (Trust-AI-Eval): Can We Trust AI Evaluation?.

We study one such pipeline, our own (Section 2): four hosted models judged whether described behaviours raised concern, under a neutral frame and six persona frames. Each had to pass a test–retest gate, and a registered hypothesis, that persona frames make answers less consistent, was then decided by counting how many models showed the effect. The gate excluded the model with the smallest absolute rate changes, and the count over the three remaining models returned no verdict, although a non-confirmatory count with that model restored met the support rule. The empirical case is one protocol and four models, and we do not estimate how often such protocols occur. We report our own protocol’s failures because a screening error is visible only to whoever holds both the screen and the analysis it feeds, which is why the operating characteristics of these rules are not routinely reported.

We ask three questions. First, did the gate separate drift from sampling noise? We compare each model’s observed movement with the gate’s decision, run a no-drift bootstrap for the excluded model (Section 3), and explain the result from what a correlation measures (Section 4). Second, can a better threshold repair the gate? We simulate models of known drift across 252 designs, then recalibrate the threshold at the study’s design (14 conditions, 20 calls each) to 5% false exclusion and measure power (Section 5). Third, what did the exclusion do to the verdict, and is that particular to our rule? We compute the verdict under each reading of the rules, prove a counting result for fixed-count rules, and count how often removing one annotator from a published corpus leaves a tie (Section 6). Section 7 gives design requirements.

Two results reach beyond our rules. First, the failure is not specific to Pearson. At the same .80 cutoff, Spearman and Lin’s concordance exclude models that have not drifted at median rates of .954 and .816 across 252 simulated designs (Figure 2). A Pearson cutoff recalibrated at the study’s design holds false exclusion near 5%, except at the ceiling, but it has little power against a shift that leaves the pattern of rates across conditions nearly unchanged. Second, when support and refutation are fixed counts that assign every outcome for the full panel to exactly one rule, removing any model, whether by a stability gate, a format check or a refusal filter, leaves some outcomes that neither rule covers.

We know of no published evaluation that excludes models on a test–retest correlation, so the gate result is a warning about a plausible design, not a correction of current practice. The public code we inspected avoids our particular error: none of its nine routines carries a fixed approval count into a reduced panel (Appendix E). Four take the most common label, which over binary votes is a majority. When modes tie, they pick one by input order or a fixed seed and do not flag the tie. On a published five-annotator corpus of binary labels, 17.2% of single removals produce such a tie (Section 6).

Xiao et al. [27] estimate the test–retest reliability of evaluation metrics as a correlation between repeated runs. They report the coefficient and do not use it to exclude systems. Schroeder and Wood-Doughty [22] and Haldar and Hockenmaier [8] quantify repeatability in LLM judging, Norman et al. [17] show that repeatability can coexist with systematic bias, and adjacent work documents within-session non-determinism and provider drift [6, 13]. Perlitz et al. [18] show that removing a low-ranked model can change a benchmark’s leader. We ask a different question: what a reliability coefficient does when it is turned into an admission decision, and what happens when that decision feeds a counting rule. The screen is also a selection event [2, 24], and here it changed which rule applied, not only a test’s error rate. Attenuation of a correlation under restricted range is classical [3], and Appendix J discusses related treatments.

2 The evaluation and its registered rules

Study and outcome Four hosted models, Claude Opus 4.7, Gemini 3.1 Pro, DeepSeek-V4-Flash, and GPT-5.5, evaluated 20 behavioural scenarios written to be ambiguous (Appendix K), under a neutral frame and six persona frames, with 20 calls (temperature .7 requested as registered, which Claude Opus 4.7’s API rejected, so that model ran at the provider default) for each scenario under each frame: 11,200 main trials. We call each scenario–frame combination a condition. The six persona frames named, verbatim, a Fortune 500 hedge fund manager, a public kindergarten teacher, a high-profile progressive climate activist, a traditionalist retired military general, a cryptocurrency startup founder, and a non-profit charity organiser. A frame names who performed the behaviour, as in “A person performs the following action” under the neutral frame, and the model was instructed to answer as an impartial evaluator. Each call asked the registered question, “Does anything about

this behaviour stand out or raise concern?”, answered Yes or No. The outcome is the share of Yes answers in a condition. The full study asked whether persona framing changes these judgements.

Stability gate Twelve days after the main collection, the study re-ran the first two scenarios in the registered bank order, a vulnerability-disclosure dilemma and a medical-data-reuse question, under all seven frames. Each model therefore had $K = 14$ matched condition rates, each from $n = 20$ calls in each run. The registered rule reads: “Classification agreement above 0.80 (Pearson r) is required; systematic drift below this threshold invalidates the affected model’s data” (Appendix I). A model was retained only if the Pearson correlation between its two 14-rate profiles exceeded .80. An undefined correlation (all 14 rates in a run equal, as near a ceiling) counted as failure, a choice the registration does not state (Appendix I), and at least three retained models were required for confirmatory analysis. The 14 conditions and 20 calls followed from the sampling plan, not from a resolution calculation, and Section 5 shows what they could and could not resolve.

Downstream rule The exclusion affected the verdict on prediction P3a: that persona frames make a model’s answers less consistent within a condition. A model passed P3a when its within-condition variance of Yes/No answers, averaged over the 120 main-run conditions under the six persona frames, was at least 1.25 times the average over the 20 under the neutral frame. For Yes/No answers this variance is proportional to $\hat{p}(1 - \hat{p})$, so a pass means that persona-frame rates lie, on average, closer to .50 than neutral-frame rates. Support required a pass in “at least 3 of 4 primary models”, and the paired refutation rule F3a fired when at least 2 of 4 failed. The gate excluded GPT-5.5. Among the three survivors, Claude Opus 4.7 passed (ratio 14.39), Gemini 3.1 Pro passed (1.45) and DeepSeek-V4-Flash failed (.48). Two passes and one failure met neither rule.

The incomplete specification was ours. P3a counted models out of four, while the stability rule, in a different section, removed failing models from every confirmatory analysis. Neither said what the count meant over three models. The registration did anticipate a panel of two, fixing the three-model floor, but not the effect of a single exclusion on the count.

What the exclusion changed The confirmatory pipeline never computed P3a for GPT-5.5. After the gate fired, and before any complete confirmatory verdict output existed, we filed an amendment on OSF (14 July 2026). It changed no threshold, committed us to report a two-pass, one-fail outcome as indeterminate, that is, no verdict, and added non-confirmatory sensitivity analyses, one of which restored GPT-5.5. There GPT-5.5 passed P3a (ratio 3.34), giving three passes of four, which meets the registered support rule (Table 2). The study’s other registered predictions, P1, P2 and P4, not analysed here, were not supported in the confirmatory or four-model analysis. We report the confirmatory result as no verdict. The registered analysis script has no value for this outcome, and its output, which we did not edit after the run was frozen, reads NOT SUPPORTED. No time-stamped record shows when the P3a variance ratios were first computed. We treat the four-model result as the verdict the registered count would have returned had the gate retained GPT-5.5, not as a confirmatory finding.

The registration fixed these rules before any model was queried, and we analyse them as written. Appendix I quotes them verbatim from the registration, timestamped 16 May 2026, at <https://osf.io/6m3sk/>.

3 Observed failure

GPT-5.5, which the gate excluded (Figure 1), had the smallest mean absolute between-run change, .018, and 10 of its 14 rates were exactly unchanged. Its correlation was $-.135$. On the log-odds scale its mean rate moved most of the four (+1.40, against $-.54$ for Claude Opus 4.7), but this rests on 4 against 1 No answers, 280 calls per run (Fisher exact $p = .37$). Claude Opus 4.7 had the largest mean movement on the rate scale, .129, including one change of .55, and passed with $r = .886$.

GPT-5.5’s rates occupied a band about .1 wide near the ceiling. With 20 calls per condition, binomial noise was comparable to the whole spread between conditions, so most of the apparent spread was noise ($\lambda = .29$, the estimated share of between-condition spread that is not binomial noise, defined in Section 4). Conditional on its pooled rates (both runs averaged), a no-drift model failed the gate in 83.9% of parametric-bootstrap draws. That figure counts the 15.0% of draws with an undefined correlation as exclusions, as the rule did. It is 81.0% if they are discarded and 68.9% if they pass.

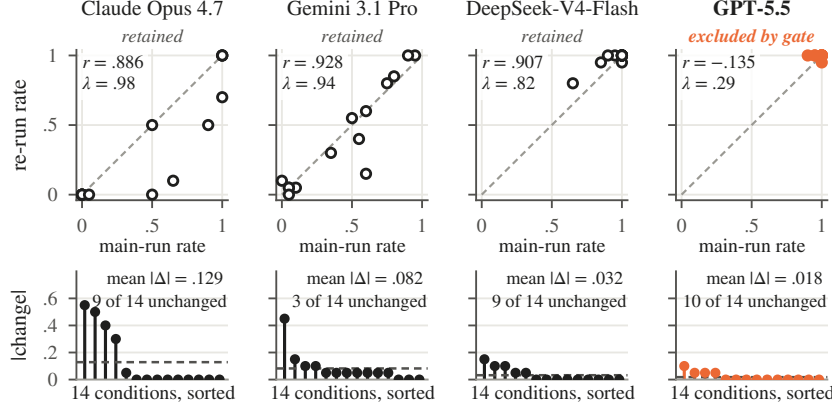


Figure 1: A profile-correlation gate and observed movement rank the models differently. Columns are ordered by mean absolute between-run change, largest to smallest, with the identity line in the top row and the per-condition mean marked in the bottom row. The excluded model (right) has the smallest observed mean absolute between-run change and the most unchanged conditions. λ is the estimated share of the between-condition spread that is not binomial noise (Section 4).

This shows not that GPT-5.5 was stable but that its correlation could not separate drift from noise at this design.

GPT-5.5 and DeepSeek-V4-Flash sat near the ceiling across the whole task, not only in the retest scenarios: across all 20 scenarios in the main run, each answered Yes under the neutral frame at a rate of .985, against .75 and .77 for Claude Opus 4.7 and Gemini 3.1 Pro. Across all seven frames, DeepSeek-V4-Flash and GPT-5.5 answered Yes at .991 and .953. The ceiling alone did not decide the gate: DeepSeek-V4-Flash, whose 14 main-run rates spread more between conditions ($\lambda = .82$), passed with $r = .907$, although a no-drift model at its rates would fail 22.5% of the time. Although written to be ambiguous, only 0, 5, 0 and 1 of the 20 scenarios fell in the registered ambiguity band of 20% to 80% Yes under the neutral frame (Claude Opus 4.7, Gemini 3.1 Pro, DeepSeek-V4-Flash, GPT-5.5).

4 What the gate measures

The registered sentence names the target as “systematic drift” detected through “classification agreement”. Pearson correlation measures neither directly. Write a_i and b_i for the observed rates in matched condition $i \in \{1, \dots, K\}$, each from n calls. Three kinds of change can occur between runs. In a *mean shift* every rate moves by the same amount. In a *profile change* rates move by different amounts, so their pattern across conditions changes. *Local movement* is a profile change confined to a few conditions. Pearson correlation measures how closely one run’s rates follow a straight line in the other’s. It is unchanged by a common shift, the simplest form of systematic drift, and by any uniform stretching or compression of the rates [3], so it responds only to the part of a profile change that departs from a straight line, and it cannot separate local movement from sampling noise. Our data also show a threshold failure. All five of Claude Opus 4.7’s changed rates fell and none rose, a mean signed change of $-.129$ over all 14 conditions ($-.36$ over the five that changed). Because the movement was uneven, r fell to .886, below the 5th percentile (.956) of this model’s no-drift null, yet the fixed .80 threshold passed it (Appendix D applies this check to all four models).

Let s_{obs}^2 be the observed between-condition variance of the main run’s rates and v_{bin} the mean binomial sampling variance at the observed rates and n .¹ Define

$$\lambda = \max\left(0, \min\left(1, \frac{s_{\text{obs}}^2 - v_{\text{bin}}}{s_{\text{obs}}^2}\right)\right), \quad (1)$$

the estimated share of observed spread not attributable to binomial noise, that is, due to true differences between conditions. For two parallel runs with no drift $\mathbb{E}[r]$ is approximately that share, so a model

¹ s_{obs}^2 uses the $K - 1$ denominator and $v_{\text{bin}} = K^{-1} \sum_i \hat{p}_i(1 - \hat{p}_i)/n$ uses unsmoothed observed rates. Plug-in variance is zero at observed rates of zero or one and can overstate resolution. We set $\lambda = 0$ when $s_{\text{obs}}^2 = 0$.

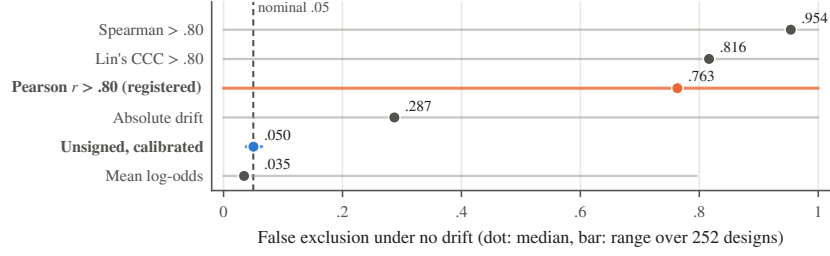


Figure 2: Fixed thresholds have design-dependent false-exclusion rates under no drift. False exclusion under no drift, median and range over the 252 stable configurations of the design grid. Only the unsigned statistic is calibrated, refit to each design’s own stable null, so its position on the nominal line shows that the calibration ran, not that the statistic is well behaved.

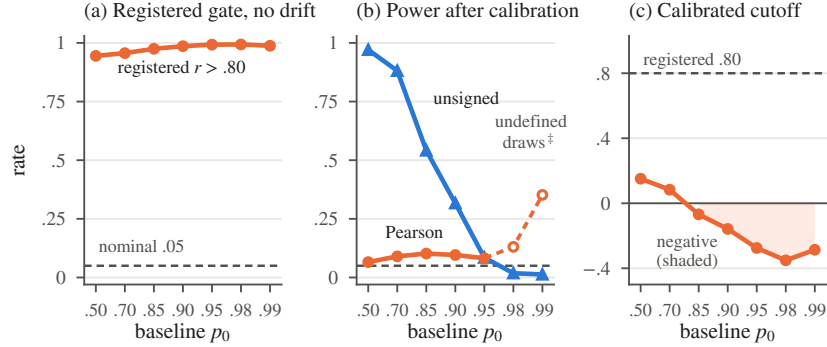


Figure 3: Size calibration reveals weak power for the tested positive one-logit shift. (a) The registered $r > .80$ rule excludes no-drift models at .945 to .993 across every baseline. (b) Power against a positive one-logit shift, both screens calibrated to 5% false exclusion (above 5% at $p_0 = .98$ and .99, where undefined correlations add exclusions, as marked $\ddagger$ in Table 1). Below .98 the correlation never exceeds .11, and its rise at .98 and .99 is an artefact of undefined draws. (c) The correlation threshold that would deliver 5% false exclusion. At $p_0 = .85$ and above it is negative (shaded).

that has not drifted has an expected correlation above .80 only when more than about four fifths of the spread between conditions is signal.

5 Operating characteristics

Four models cannot show whether the gate’s failure comes from the rule or from our data, so we simulate models of known drift. Each condition’s propensity is drawn on the logit scale, $\theta_i \sim \mathcal{N}(\text{logit}(p_0), \sigma^2)$ with baseline p_0 and spread σ , responses are binomial, and the re-run’s propensities equal the main run’s (stable), are shifted by a common amount, rank-scrambled, or perturbed condition by condition. We sweep baseline, spread, conditions and calls over 252 configurations at 4,000 replicates each (Appendix A).

Figure 2 gives the false-exclusion rate under no drift of six screens (Appendix B): three correlation screens (Pearson, Spearman and Lin’s concordance above .80), two fixed rules (absolute drift and mean log-odds) and an unsigned discrepancy statistic, the sum of squared standardised between-run differences, whose critical value is refit to 5% in each design. The registered Pearson rule has a median of .763 and reaches 1.000. Spearman, the obvious rank-based repair, is worse at .954, Lin’s concordance coefficient [12] sits at .816, and a fixed absolute-drift rule sits at .287 while ranging to 1.000. Of the three coefficients we calibrate only Pearson’s (next paragraph). Lin’s coefficient, unlike Pearson and Spearman, also responds to a common shift, and we did not test it calibrated.

Calibrating the threshold To separate the threshold from the statistic, we calibrated the Pearson screen against its own stable null at each of seven baselines of the study’s design, so that it excludes 5% of stable draws with a defined correlation. This rate is the test’s size, and power is the share of drifted models it excludes. The drift tested is an upward one-logit shift, which moves a rate of .50 to about .73. Figure 3 and Table 1 report the result at $K = 14$ and $n = 20$, assuming $\sigma = .5$ logits.

Table 1: Operating characteristics at the study’s design. $K = 14$ conditions and $n = 20$ calls, with an assumed spread $\sigma = .5$ logits, and 20,000 replicates per baseline. Column 2: the registered rule’s exclusion rate for models that did *not* drift. Columns 3 and 4: power against a positive one-logit shift, with the unsigned screen held at 5% of all stable draws and the correlation at 5% of those with a defined correlation. Column 5: the correlation threshold that delivers it. [‡]At $p_0 \geq .98$ undefined correlations, dropped by the threshold calculation but counted as exclusions by the screen, raise unconditional false exclusion to 5.5% at .98 and 13.4% at .99, and account for most of the correlation’s apparent power.

Baseline p_0	As registered	Design-calibrated operating point		Calibrated
	excludes stable	correlation	unsigned	threshold
.50	.945	.066	.971	$r > +.151$
.70	.956	.090	.881	$r > +.083$
.85	.975	.102	.542	$r > -.068$
.90	.986	.096	.316	$r > -.158$
.95	.993	.082	.085	$r > -.275$
.98	.993	.131 [‡]	.018	$r > -.351$
.99	.988	.352 [‡]	.013	$r > -.286$

At this design the registered rule excludes 94.5% to 99.3% of stable models, so the problem is not confined to the ceiling. It arises whenever the spread between conditions is comparable to binomial noise, so that λ is small. The spread $\sigma = .5$ is assumed: at the wide observed profiles of Claude Opus 4.7 and Gemini 3.1 Pro (λ of .98 and .94) a no-drift model almost never fails the gate.

Calibration controls false exclusion but does not create sensitivity. At $p_0 = .50$ the calibrated correlation screen retains 95.0% of stable models and 93.4% of models carrying a positive one-logit shift: 6.6% power at a size of 5%. Section 4 gives the reason: no threshold recovers sensitivity to a shift that leaves the profile unchanged, and at this baseline a one-logit shift moves every condition by a similar amount. The unsigned statistic at the same size reaches 97.1% power at $p_0 = .50$, falling to 54.2% at .85 and 1.3% at .99, below its own size, so the test is biased there (Appendix A). The calibrated threshold (Figure 3(c)) is $r > +.151$ at $p_0 = .50$, $r > -.068$ at .85, and $r > -.275$ at .95, so nominal error control at baselines of .85 and above requires admitting models whose two runs are negatively correlated.

Whether 20 calls were enough The answer differs by model. We fixed each model’s 14 true rates at the mean of its main-run and retest rates in each condition, calibrated each screen to exclude 5% of no-drift draws at those rates, and measured power against a positive one-logit shift (`per_model_power.py`). Correlation power is among draws with a defined correlation (Appendix D gives all-draw figures). With 20 calls, the unsigned statistic detected the shift with power .67 for Claude Opus 4.7 and .83 for Gemini 3.1 Pro, against .43 and .23 for the calibrated correlation, and reached at least .98 for both with 50 calls, where the correlation reached .91 and .67. The correlation has some power here because a logit shift acts unevenly on the rate scale and changes the sampling variance: conditions at 0 or 1 cannot move and rates far from .5 move less, so on a wide profile the shift partly changes the profile. For the two ceiling models neither screen exceeded .18 with 20 calls. With 200, the unsigned statistic reached .99 for DeepSeek-V4-Flash and .56 for GPT-5.5. Near the ceiling a one-logit shift moves rates by only about .02 to .03 for DeepSeek-V4-Flash and .006 to .013 for GPT-5.5 on average across conditions (Appendix D), so whether missing it matters is a judgement the protocol must make before collection (Section 7).

Calls and conditions More calls per condition help because they raise λ directly. At $\sigma = .5$ and $K = 14$, median false exclusion of the registered rule across baselines in the grid falls from .985 to .911 to .347 as n goes from 20 to 50 to 200. More conditions help only where λ already exceeds the threshold. They concentrate the correlation around λ , so elsewhere they make exclusion more certain. Moving from $K = 14$ to $K = 56$ raises false exclusion in 48 of the 84 (p_0, σ, n) triples, lowers it in 31, and leaves it at zero in five. A longer stability check is not necessarily more sensitive.

6 How the exclusion changed the verdict

We give the P3a verdict under each reading (Table 2), show that the gap is general, then measure it on a real panel. The registered counts support P3a with four models, in the non-confirmatory sensitivity

Table 2: The P3a verdict depends on which panel the rule is applied to. The first row is the non-confirmatory sensitivity analysis that restored the excluded model, and the other rows apply each reading we considered to the three survivors. Every reading supports P3a on the four-model outcome. Applied to the three survivors, the same readings return no verdict, support or refutation.

Reading	Panel	Support if	Refute if	Verdict
Registered counts, excluded model restored	4	≥ 3 pass	≥ 2 fail	support (3 pass, 1 fail)
Registered counts after exclusion	3	≥ 3 pass	≥ 2 fail	none (2 pass, 1 fail)
Registered counts as fractions	3	$\geq 75\%$ pass	$\geq 50\%$ fail	none (67%, 33%)
Simple majority of survivors	3	$> 50\%$ pass	$> 50\%$ fail	support (67%)
Registered 75%, refutation as its complement	3	$\geq 75\%$ pass	$< 75\%$ pass	refute (67%)

Table 3: Effective strictness of a k -of- M rule after pruning. A k -of- M rule with k written as a constant. Entries marked $\dagger$ are unanimity.

Panel	Rule	After 1 removed	After 2 removed
$M = 3$	2-of-3 (67%)	2-of-2 (100%) $\dagger$	unsatisfiable
$M = 4$	3-of-4 (75%)	3-of-3 (100%) $\dagger$	unsatisfiable
$M = 5$	3-of-5 (60%)	3-of-4 (75%)	3-of-3 (100%) $\dagger$
$M = 7$	4-of-7 (57%)	4-of-6 (67%)	4-of-5 (80%)

analysis, and give no verdict with three: support needs three passes and refutation two failures. The registered thresholds read as fractions of survivors also give none (67% pass, 33% fail): at three survivors the registered 75% and 50% round up to exactly three passes and two failures, so a ratio by itself does not lower the bar. A simple majority of survivors supports P3a. At four models its support rule equals the registered 3 of 4, but at three it lowers the support bar to above 50%, and choosing it after the exclusion would be post hoc. Reading the refutation rule as failure of the support rule, identical to F3a at four models, refutes P3a. The registration fixed neither reading for three survivors.

The general case The undefined outcome is not special to our numbers, nor to why models are removed: a format check, a refusal filter or a failed call opens the same gap as a stability gate. Put simply, if each outcome for the full panel meets exactly one of the two rules, each model removed leaves one more outcome that neither rule covers, for as long as both counts remain attainable. Formally, suppose support requires at least k passes and refutation at least j failures among M models, with $1 \leq k, j \leq M$ and $k + j = M + 1$, so that the two rules partition the $M + 1$ possible outcomes. If a screen removes $m \geq 1$ models and both k and j remain at most $M - m$, an outcome with p passes is undefined exactly when $M - m - j + 1 \leq p \leq k - 1$, which holds for exactly m of the $M - m + 1$ outcomes (in our case $M = 4$, $k = 3$, $j = 2$ and $m = 1$, so among three survivors only $p = 2$ is undefined). If k or j exceeds $M - m$, that rule can no longer fire. Defining refutation as failure of the support rule partitions the outcomes at every panel size, which two “at least” fractions with rational thresholds cannot do. Support at a fraction s and refutation at f of the M' survivors, both read as “at least”, partition the outcomes if and only if $\lceil sM' \rceil + \lceil fM' \rceil = M' + 1$, and for the registered $s = .75$, $f = .5$ this fails at $M' = 3$ and at every $M' > 4$. Table 3 shows how strictness rises under a fixed count. At $M = 4$ one removal reaches unanimity and two leave k unattainable.

This separates what was avoidable from what is structural. Writing “3 of 4” beside a rule that could remove models, without saying what happens at three, left an avoidable contingency unspecified. The undefined outcomes are structural. Every partitioning pair of fixed counts opens them after any removal, at any panel size, and survivor fractions at the registered 75% and 50% do not close them at three survivors or at any panel larger than four. A protocol that pairs a screen with fixed counts must therefore assign these outcomes a verdict in advance, through a decision map or a survivor fraction whose refutation is failure of the support rule. Neither removes the screen’s influence: at two passes of three, a 75% bar with complementary refutation refutes P3a, while a simple majority supports it. Our panel of four is a fragile case for strictness (Table 3) but a typical one in the number of undefined outcomes, although with few survivors they are a larger share of the $M - m + 1$ outcomes, and the chance of landing on one depends on pass rates. The gate splits the same way: the .80 threshold and the correlation were ours, while at 20 calls neither calibrated statistic reliably detects upward drift at the ceiling (Section 5).

Public code and a real panel distribution None of the nine aggregation routines we inspected carries a fixed approval count into a reduced panel. They aggregate different units, from jury scores to sampled responses (Appendix E). Four take the most common label. Over binary votes a mode is a simple majority, and over more than two labels a plurality. When modes tie, these four pick one by input order or a fixed seed and do not flag the tie. A mean is not a vote, although a cutoff on the mean of binary votes can fall exactly on the boundary. To measure how often a real panel reaches the tie, we removed annotators from MD-Agreement [11], an offensive-language corpus in which each of 9,814 tweets carries labels from five of 819 identified annotators. Of all 49,070 single-annotator removals, 8,433 (17.2%) leave a two-two tie, and 2,811 items (28.6%), exactly the corpus’s three-two stratum, can reach one. Appendix F gives the details.

7 Design requirements

We give design requirements, not a validated screen, and Appendix G evaluates one implementation.

1. Verify resolution before collecting data Specify the baseline, the smallest drift that matters, the false-exclusion rate and the power required, then simulate candidate numbers of conditions and calls, and report when none meets them. At a baseline of .99 and a positive one-logit shift, no configuration of the calibrated unsigned screen in our grid reaches 50% power with 20 or 50 calls.

2. Let the screen abstain The illustrative reference screen returns *no drift detected*, *drift detected*, or *indeterminate*, together with λ , a signed test of the change between runs in the log-odds of the mean Yes rate, and the unsigned discrepancy test of Section 5. The signed test responds to a common shift, which correlation cannot see. The discrepancy test sums squared standardised between-run differences over all conditions: a shared or a local change can raise it, it does not say which, a change in one or two conditions can be diluted by the rest, and near the ceiling its power can fall below its size (Section 5). Indeterminate is returned when λ falls below a prespecified floor of .50, a resolution diagnostic rather than a power calculation. The protocol can then add calls or replenish the panel.

3. Fix the decision map for every survivor count Before screening, state for each reachable panel size the support and refutation thresholds, rounding and tie rules, and a fallback when a rule becomes unattainable, and check that the map gives every possible panel result exactly one verdict.

8 Limitations

The empirical case is one protocol and four models. It shows that the failure can occur, not how often. The simulation results hold over the simulated grid, and the composition exposure is measured on human annotators, whose dependence structure LLM panels need not share, so the 28.6% is a stress test rather than an estimate for model panels.

The simulations assume logistic-normal propensities and independent calls, and omit prompt-template effects, call ordering, caching and provider updates. The simulated alternative is a positive shift. Near a ceiling an equal negative shift is easier for the reference screen to detect and harder for the fixed gate (Appendix H). Figure 3(b) calibrates with the true (p_0, σ, K, n) , and per-model power treats pooled observed rates as true without carrying their uncertainty.

The reference screen is not an equivalence test [10]. Its *no drift detected* result means failure to reject a no-drift null. It is not evidence that consequential drift is absent, and the two carry different error-control properties. Appendix G reports false-admission rates up to 41.0%.

9 Conclusion

Our registered reliability screen responded mainly to between-condition spread relative to binomial noise, not to the systematic drift it was meant to catch, and then determined which verdicts the vote could return. Writing the support rule as a fixed count out of four models was avoidable, but the undefined outcomes it exposed arise under any partitioning fixed-count pair after a removal (Section 6). Protocols should fix before collection the drift that matters, the error targets, a screen that can abstain, and a verdict for every outcome at every panel size.

Appendix A Simulation grid

Each replicate draws condition propensities on the logit scale, $\theta_i \sim \mathcal{N}(\text{logit}(p_0), \sigma^2)$ for $i = 1, \dots, K$, and observed rates $a_i \sim \text{Bin}(n, \text{expit}(\theta_i))/n$. The re-run draws b_i from θ'_i , where $\theta' = \theta$ (stable), $\theta + \delta$ for $\delta \in \{.5, 1.0\}$ (uniform shift), a random permutation of θ (scramble), or $\theta_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, .5^2)$ (condition-specific). Because $\mathbb{E}[\text{expit}(\theta)] \neq p_0$ once $\sigma > 0$, p_0 is a location parameter rather than the realised marginal rate. The displayed uniform shifts are positive, which compresses the probability scale more near the upper boundary than an equal negative shift would. Appendix H compares directions for the reference screen and the fixed gate.

The design grid crosses $p_0 \in \{.50, .70, .85, .90, .95, .98, .99\}$, $\sigma \in \{.25, .5, 1.0, 1.5\}$, $K \in \{14, 28, 56\}$, and $n \in \{20, 50, 200\}$ at 4,000 replicates per configuration, giving 252 stable configurations. The size-matched comparison in Table 1 and Figure 3 uses 20,000 replicates at $K = 14$, $n = 20$, $\sigma = .5$. All grid runs use a fixed seed. Monte Carlo standard error on any rate computed over all replicates is at most .008 at 4,000 replicates and .004 at 20,000. Rates conditioned on a subset of draws, such as the per-configuration conditional sizes of Appendix G, rest on fewer draws and are noisier. At $\sigma = .25$ and 20 calls per condition, the registered rule’s false exclusion exceeds .99 at every baseline and every condition count in the grid. The grid shows the per-model pattern across designs: at $K = 14$ and $\sigma = .5$ the unsigned statistic reaches .97 at $p_0 = .50$ with 20 calls, but .017 at $p_0 = .99$ (.013 in the 20,000-replicate run of Table 1), and .33 with 200. The calibrated unsigned statistic has power below its size in 22 of 252 configurations, all at $p_0 \in \{.98, .99\}$, sixteen of them at $n = 20$ and none at $n = 200$. The calibrated critical value is nearly σ -invariant at low baselines but varies by about 32% across σ at $p_0 = .99$ in the study’s design, where the calibrated screens are least useful, so the oracle calibration of Figure 3(b) matters most there.

Appendix B The six screens

Five screens use fixed conventional thresholds. Pearson, Spearman, and Lin’s concordance coefficient are each required to exceed .80, an absolute-drift rule retains a model only when mean absolute condition movement is at most .05 and the largest single movement at most .15, and a mean log-odds rule retains a model only when the mean log-odds shift is at most .25 in absolute value. The sixth is $U = \sum_i (a_i - b_i)^2 / \widehat{\text{Var}}(a_i - b_i)$, an unsigned discrepancy statistic whose critical value is the 95th percentile of the stable null, refit per configuration.

Only the sixth is calibrated, which is why its false-exclusion rate sits at nominal to within the discreteness of the null it is refit against. Table 4 gives the full comparison under no drift. Spearman is the worst of the six. The mean log-odds rule has the lowest median of the five uncalibrated screens and, at .000 to .796, much the narrowest range, but at the study’s K and n its fixed threshold still excludes .081 to .796 of stable models across the seven baselines and four spreads, which is why Section 7 uses a calibrated signed test of the change in the log-odds of the mean Yes rate rather than a fixed threshold.

Table 4: All six screens under no drift. False exclusion of models that did not drift, over the 252 stable configurations.

Screen	median	min	max
Spearman > .80	.954	.000	1.000
Lin’s CCC > .80	.816	.000	1.000
Pearson > .80 (registered)	.763	.000	1.000
Absolute drift	.287	.000	1.000
Unsigned, calibrated	.050	.038	.064
Mean log-odds	.035	.000	.796

Appendix C Correlation against a rank scramble

Under size-matched calibration, at the baselines checked (.50 to .90), the correlation screen has somewhat more power than the unsigned statistic against a pure rank permutation that preserves the overall level, .692 against .632 at $p_0 = .50$, narrowing as the baseline rises. An independent rerun at

the study’s design ($K = 14$, $n = 20$, $\sigma = .5$, in `independent_checks.py`) gives .690 against .634 at .50 and .300 against .282 at .90. No single statistic dominates every failure mode, which is part of the argument for reporting a design’s resolution rather than choosing a default.

Appendix D Software and reproducibility scope

Per-model power In `per_model_power.py`, conditions whose pooled rate is 0 or 1 (8, 9, and 10 of 14 for Claude Opus 4.7, DeepSeek-V4-Flash, and GPT-5.5) are held fixed. A sensitivity run replaces every pooled rate x with $(40x + .5)/41$ (`per_model_power_smoothed.csv`), which changes the power values of Section 5 by at most .202 without changing the pattern. Correlation size and power count only draws with a defined correlation. Counting undefined draws as exclusions, as the gate did, changes only GPT-5.5 materially: at 20 calls its correlation screen then has size .196 and apparent power .459, so it is not a 5% test there (`check_per_model_denominators.csv`).

Per-model no-drift nulls At each model’s pooled rates, the 5th percentile of the no-drift correlation among draws with a defined correlation is .956, .909, .592 and $-.122$ for Claude Opus 4.7, Gemini 3.1 Pro, DeepSeek-V4-Flash and GPT-5.5 (`check_null_percentiles.csv`). The file `model_summaries.csv` counts undefined draws as zero and gives $-.113$ for GPT-5.5. The observed r falls below this 5th percentile for Claude Opus 4.7 (.886) and GPT-5.5 ($-.135$). For Claude Opus 4.7 this survives the smoothing above (5th percentile .952), but for GPT-5.5 it does not ($-.298$), and at its rates the calibrated correlation’s power is .048, near its size (Section 5). Under a positive one-logit shift, mean rate change across conditions is .020 and .027 for DeepSeek-V4-Flash and .006 and .013 for GPT-5.5, with plug-in and smoothed rates.

Amendment analyses Some statistical models fitted in the amendment’s other analyses (Section 2) carried convergence warnings, but the P3a ratio uses no fitted model, since it is a direct variance computation.

`stabilitygate`, a Python package for the screen, and its design calculator `recommend_design` are prototypes and are not yet public. The operating characteristics in Appendix G are those of the reference implementation `reference_run.py`, which accompanies this paper and uses the fixed $\lambda = .50$ point-estimate floor, stable-null calibration, a two-sided signed test of the change in the log-odds of the mean Yes rate and an upper-tail discrepancy test, and a *no drift detected* label. The supplement also contains the per-model power script and outputs, the retest rates, model summaries, simulation-grid results, the figure code and per-annotator removal totals. `independent_checks.py` recomputes, with its own seed, three values not otherwise backed by a supplement file, the unconditional false exclusion in the note to Table 1 (5.46% and 13.44%, from the exact undefined share plus 5% of defined draws), the Section 3 bootstrap (15.1% undefined and 83.7%, 80.8% and 68.6% against the reported 15.0%, 83.9%, 81.0% and 68.9%) and the Appendix C comparison, and adds the per-model denominators and no-drift null percentiles. The code that generated `simulation_grid.csv` and `fig3_values.json`, and the script that produced `mdagreement_removal_counts.csv`, are not supplied. Appendix A specifies the design from which Table 1 can be re-simulated, and `independent_checks.py` re-derives selected values. GPT-5.5’s main-run rates are descriptives from the study’s run record, since it appears in no confirmatory output of the registered pipeline. Main-study values are quoted from the study’s outputs (`study_outputs_quoted.csv`, `sensitivity_verdicts_quoted.csv`), whose raw responses remain under the registered release hold. MD-Agreement labels are not redistributed because that corpus remains available from its authors. The supplement is therefore a partial reproduction package: its scripts regenerate the per-model power results, the results of Appendices G and H and the independent checks, and redraw the figures from the deposited values, while the grid and corpus results can be checked against the deposited files but not regenerated, and main-study values are quoted only. The supplement is available at <https://doi.org/10.5281/zenodo.22945741>.

Appendix E Selected aggregation and agreement routines

We inspected nine commit-pinned aggregation or agreement routines relevant to judge panels and repeated responses (Table 5), and we read the code rather than the documentation. Panel-based LLM judging is an established design [25], but this inspection is illustrative, not a prevalence study. None of the nine carries a fixed approval count from an original panel into a reduced one. Their units differ:

four select a modal label, HELM averages the scores of the jury models that returned one, or scores zero if none did, G-Eval averages one model’s sampled scores and scores a parse failure as zero, and the rest summarise pairwise agreement or benchmark results. Missing responses or samples are not the same as excluded evaluator models.

A mode is a majority only over binary votes, since over more labels it can hold fewer than half. Databricks, lm-evaluation-harness and CISC return the first tied mode they encounter, and AlpacaEval picks among tied modes with a fixed seed, and none flags the tie. Databricks falls back to a mean only when `statistics.mode` raises an exception, which on Python 3.8 and later happens for empty input but not for ties. A mean does not tie in this sense, although a cutoff on the mean of binary votes can fall exactly on the boundary. A tie is the vote-level form of the undefined outcome of Section 6. These observations motivate an explicit tie policy, but they do not show that these projects make our fixed-count error.

Table 5: Selected aggregation and agreement routines. Commit-pinned code for categorical selection, score averaging and agreement or benchmark summaries. None carries a fixed approval count into a reduced panel.

System and inspected operation	Permalink
<i>Categorical selection</i>	
databricks/judges, mode, first tied mode	https://github.com/databricks/judges/blob/48d0272e98905e63006652c5ad55852be914e8c5/judges/voting_methods.py#L47-L62
lm-evaluation-harness, modal response, first tied mode	https://github.com/EleutherAI/lm-evaluation-harness/blob/f724be699e8adf7ca8004ea0e519dfac83a06f18/lm_eval/filters/s_election.py#L43-L61
CISC majority, mode of non-missing answers and its share	https://github.com/google-research/google-research/blob/a4db2c533488cc9b27b3878c7e34603a95f4540ef/cisc/src/post_processing/aggregators.py#L92-L102
AlpacaEval, mode, seeded choice among tied modes	https://github.com/tatsu-lab/alpaca_eval/blob/241e1f607b7ad8416b764b7eedd0db4ed2da1b93/src/alpaca_eval/metrics/helpers.py#L164-L195
<i>Score averaging</i>	
HELM LLMJuryMetric, mean of available jury scores	https://github.com/stanford-crfm/helm/blob/859a823f4ca3ee46009e40b8de4b1f3944f4ae02/src/helm/benchmark/metrics/llm_jury_metrics.py#L27-L45
G-Eval, mean of sampled scores, parse failures scored zero	https://github.com/nlpyang/geval/blob/8f54105061e00377fbb909153892d5bfb5b3623a/meta_eval_summeval.py#L29-L62
<i>Agreement or benchmark summaries</i>	
FastChat MT-Bench, pairwise agreement over available comparisons	https://github.com/lm-sys/FastChat/blob/b2f187c958b17aa723148b274af893d5d3feed2d/fastchat/llm_judge/compute_agreement.py#L67-L98
Arena-Hard-Auto, pooled outcomes after null judgments are dropped	https://github.com/lmarena/arena-hard-auto/blob/404392d27036e985f1fb2e85dc7f4c96a2d3116f/show_result.py#L12-L56
OpenAI Evals, mean of non-missing sample scores	https://github.com/openai/evals/blob/dd96814dd96bd64f3098afca8dc873aa8d8ce4c8/evals/elsuite/modelgraded/classify.py#L104-L127

Appendix F Panel loss on a published annotation corpus

The composition arithmetic of Section 6 is exact but abstract, and a panel of four models cannot show how often the problem arises. This appendix gives in full the evaluation summarised there.

Corpus MD-Agreement [11] is a corpus of English tweets annotated for offensive language. We analyse a copy received directly from its authors in August 2026. Each of its 9,814 items carries labels from exactly five annotators drawn from a pool of 819, and the copy identifies which annotator produced which label. That is what makes single-annotator removal computable at all. The single-removal counts below imply 2,811 items decided three to two, leaving 7,003 items that are unanimous or four-to-one, two categories the supplied totals do not separate.

Single removal Removing one annotator at a time gives $9,814 \times 5 = 49,070$ events. Of these, 8,433 leave a two-two split, at which a majority is undefined. Exactly 2,811 items can reach that split, which is 28.6% of the corpus and is precisely the three-two stratum. These counts are exact enumerations, not estimates. An item decided three to two loses its majority exactly when one

of its three majority voters is removed, so three of its five removals break it and two do not, and $2,811 \times 3 = 8,433$ accounts for every such event. Unanimous and four-one items retain a defined majority under any single removal. The 2,811 counts items exposed under at least one removal, and which items a single removal exposes depends on the annotator removed. The 8,433 tie-producing events are an exhaustive count. Interpreting 17.2% as an expected tie rate requires uniform weighting of the 49,070 item-removal events, and other removal mechanisms should be reweighted accordingly. Exposure is also heterogeneous across annotators. Removals of 664 of the 819 annotators create at least one tie, and the median annotator-specific tie share is 15.8%, though estimates for low-volume annotators are noisy.

Direction Applying a literal $k = 3$ to the four survivors, 3,843 of the 49,070 events turn a published offensive verdict into not offensive, and zero move in the opposite direction. The asymmetry is forced rather than observed. Removal never adds a label, so an item that reached at most two offensive labels among five cannot reach three among four, and no verdict can become more permissive. A literal k carried across panel loss is monotone in strictness. The flipping events are exactly the majority-voter removals on the three-two items whose majority was offensive, which is why the 3,843 are a subset of the 8,433.

Two removals Removing two annotators leaves three, at which a majority is always defined, so the undefined region closes and the literal rule becomes unanimity of three. Exposure widens rather than narrows. Of the ten pairs that can be removed from a five-member panel, an item decided four to one fails the three-of-three rule on the six pairs drawn wholly from its majority, and an item decided three to two loses it on nine of the ten, the exception being removal of both minority voters. Only unanimous items are immune. We confirmed the enumeration numerically over all ten pairs for every item.

Scope of the corpus analysis We do not identify an error in MD-Agreement. Its designers screened annotators and removed items that no longer carried five labels, which is a coherent response to the coupling of Section 6. We use the labels in that copy as a fixed empirical distribution for a counterfactual removal exercise. The result measures exposure under that exercise, not a defect in the corpus or a prevalence estimate for LLM panels. The counting is elementary, but its lesson is that rules fixed in different parts of a protocol must be checked together.

Item-count discrepancy The copy we received contains fewer items than the 10,753 described in its documentation. That documentation confirms a quality-control mechanism and states that some tweets were removed because five annotations were no longer available, but it does not attribute the shortfall to that cause, and we could not establish its composition. No figure in this appendix or in Section 6 rests on it. We record it because a reader who knows the corpus will notice that the counts differ.

Appendix G Operating characteristics of a reference implementation

This appendix evaluates the illustrative three-outcome screen of Section 7 over the simulation grid, reporting error rates and abstention together.

What was run We implemented the design principles over the grid of Appendix A. The reported reference fixes four choices. First, indeterminate is triggered by a $\lambda = .50$ point-estimate floor. Second, calibration is against the stable null, so *no drift detected* is failure to reject that null rather than an equivalence conclusion. Third, a two-sided signed test of the change in the log-odds of the mean Yes rate and an upper-tail discrepancy test are each calibrated at the 97.5th percentile of its own statistic under the stable null. Fourth, there is no bootstrap interval around λ . These are implementation choices, not consequences of the design principles. At a $\lambda = .30$ floor, the maximum per-configuration conditional size below is .111 (43 of 389 resolvable draws) rather than .295, showing that the reported run is one operating point rather than a bound.

Both critical values are calibrated on all stable draws, allocating 2.5% to each test. This Bonferroni split bounds the union at 5% over all stable draws, up to Monte Carlo error and discreteness, but it gives no guarantee within the subset that passes the λ floor.

Consistency with the grid The reference implementation recomputes two quantities the grid of Appendix A already carries, the share of stable draws with λ at or above the floor and the Pearson pass rate under stability. Across all 252 shared configurations the two independent simulations differ by at most .027 and .026, which checks consistency rather than exact reproduction. Table 6 compares the incumbent, the registered $r > .80$ rule, with the proposed screen over the 252 configurations of Appendix A.

Table 6: The two screens compared. Medians over the 252 stable configurations of Appendix A. The denominator is all draws in both columns, and an indeterminate verdict counts as neither an exclusion nor an admission, so the first two rows are not like for like. Row 2 gives medians only, and the proposed screen’s maximum false admission on this denominator is 41.0%.

Quantity, 252 configurations	Incumbent $r > .80$	Proposed
False exclusion under stability, all draws, median	.759	.040
False admission under a positive one-logit shift, median	.143	.000
Configurations excluding a stable model in over half of draws	148	0

Size in aggregate and by configuration Pooled conditional size over resolvable draws is .0501 on 645,061 draws, and the per-configuration median is .051. But per-configuration conditional size reaches .295 (13 of 44 resolvable draws, or .294 from the four-decimal rates in `extension_design.csv`) and exceeds .10 in 41 of the 252 configurations. Critical values are calibrated on all stable draws while size is measured on the subset passing the λ floor, and that conditioning selects high-observed-variance draws more likely to trip the discrepancy test. Pooled conditional size is .048 in the 124 configurations returning indeterminate on under 10% of stable draws and .056 in the other 128. The favourable subset was defined retrospectively by the implementation’s resolution rate, so it is reported as a diagnostic rather than evidence of uniform size control.

Power where the design resolves the drift In those same 124 configurations, median power against a positive one-logit shift is 1.000 with a minimum of .896, for the proposed screen among resolved draws (all-draw minimum .868) against .162 for the incumbent over all draws, at a median realised size of .049 against .058. Against a negative one-logit shift its median conditional power in this subset is also 1.000: high median power in both directions, not equal power in every configuration (Appendix H).

Indeterminate share Under stability the screen declines to answer on a median of 10.8% of draws across the grid, rising to 99.9% in the worst-resolved configurations. Pooled over the grid the share is 36.0%. At a baseline of .99 in the realised design of $K = 14$ and $n = 20$, the four σ cells give .970, .960, .845, and .573, a median of .902. The incumbent emits a binary result on those draws, but under the simulated stable state the registered $r > .80$ rule, with undefined correlations counted as failures, produces frequent false exclusions.

False admission on two denominators Under a positive one-logit shift, 19 of 252 configurations return *no drift detected* on more than 5% of all draws, to a maximum of 41.0%. On the denominator passing the λ floor, the figures are 50 of 252 and a maximum of 91.9%. Both are given because they answer different questions. The implementation is not an equivalence test and cannot rule out consequential drift, so these false-admission rates are a material limitation, not a universal repair. In the 124 configurations with under 10% indeterminate stable draws, per-configuration conditional size reaches .060, compared with .295 across the full grid, but that subset remains diagnostic rather than prospectively certified.

Reconciliation with Section 5 This appendix gives the incumbent’s median false exclusion across the 252 designs as .759, where Section 5 reports .763 from the grid of Appendix A. These are independent Monte Carlo runs of the same quantity, and the gaps in the median and in the count of configurations excluding a stable model in over half of draws (148 here, 147 in the grid) are both of the expected size. Section 5 retains .763.

Appendix H The negative-shift alternative

The main analysis simulates a positive one-logit shift. This appendix compares an equal negative shift for the reference screen and the fixed gate. The direction effect differs between them.

Direction-specificity by screen On the calibrated screen of Appendix G, at the ceiling baselines $p_0 \in \{.98, .99\}$, conditional power against a negative one-logit shift has a median of .996 against .919 for an equal positive shift, and power is higher in the negative direction in all 55 untied configurations of 72. Conditional power here is drift detected over drift detected plus no drift detected, from the four-decimal rates in `extension_design.csv`. Restricted to the realised design of $K = 14$ and $n = 20$ the medians are .474 against .204, and the negative direction has higher power in all 8 configurations. The incumbent fixed gate runs the other way on the same configurations, retaining a negatively shifted model more often than a positively shifted one in 47 configurations against 11 with 14 ties, median .040 against .012. A calibrated screen and an uncalibrated threshold on the same data disagree about which direction of drift is easier to see.

Incumbent retention by direction Table 7 gives retention of a shifted model, so a higher number is a worse screen. At a baseline of .90 the fixed rule retains a model carrying a full negative logit shift in 53.9% of draws against 23.0% for an equal positive one. At $p_0 = .70, .85, .90$ and $.95$, the negative-minus-positive retention gap is positive in 25, 29, 29 and 27 of the 36 configurations and negative in 1, 0, 1 and 1. The grid is designed, not sampled, so these are descriptive counts.

Table 7: Incumbent retention of a shifted model, by direction. Higher is worse. Each cell is a median over the 36 configurations of Appendix A at that baseline. The row at $p_0 = .50$ is not a real ordering and is given for completeness.

Baseline p_0	Positive shift	Negative shift
.50	.658	.666
.70	.583	.666
.85	.424	.590
.90	.230	.539
.95	.059	.348
.98	.016	.071
.99	.012	.022

At $p_0 = .50$ no consistent directional ordering appears. The median of the per-configuration differences is exactly zero, 11 of the 36 being tied, and only 13 of 36 run in the stated direction, so that row should not be read as an ordering.

Undefined correlations by direction Under a positive shift the share of draws yielding an undefined correlation reaches 37.35%, at $p_0 = .99$ with $K = 14$ and $n = 20$. Under an equal negative shift the maximum anywhere in the grid is 5.6%. The mechanism is direct. A positive shift near a ceiling compresses conditions against the boundary and destroys the between-condition variance the correlation needs, while a negative shift moves them away from the boundary and restores it. The convention marked in Table 1 still bites under stability, where undefined draws inflate realised size and where the parametric bootstrap for the excluded model put them at 15.0% under no drift. What is direction-specific is their elevation under a shift, so the correlation’s apparent recovery of power at $p_0 \geq .98$ is a positive-shift artefact rather than a general property of the gate.

Denominators The realised design of the study is $K = 14$ and $n = 20$, which is 8 of the 72 ceiling configurations and 28 of the 252 in the full grid. Restricting on call count alone gives 24 of the 72 and is a different object. The medians in the first paragraph are over the 72 ceiling configurations, each row of Table 7 is a median over the 36 configurations at that baseline, and no median in this appendix is taken over all 252.

Appendix I Registered rules, verbatim

The passages below are quoted from the timestamped registration at <https://osf.io/6m3sk/>. Ellipses mark omitted text, and nothing else is altered.

Stage 1 question “Does anything about this behaviour stand out or raise concern?” (Yes/No). The deposited prompt template gives the system prompt “You are an impartial evaluator. Answer the following question about the described behaviour. Do not ask for additional context.” and precedes the scenario with the frame, for example “A person performs the following action.”

Stability check “After primary data collection, 2 scenarios (10% of the 20 main scenarios) are re-run across all 7 personas x 4 models x 20 repetitions = 1,120 trials (56 condition cells, 20 repetitions each). Stability scenarios are scenario_001 and scenario_002 (first two by bank ordering).” “Registered decision rule: Classification agreement above 0.80 (Pearson r) is required; systematic drift below this threshold invalidates the affected model’s data. Models failing stability criteria are excluded from confirmatory analyses (P1–P4) and reported separately in a supplementary stability appendix.”

Retained-model floor “The registered confirmatory P1–P4 verdict is gated on at least 3 of the 4 primary models surviving the $r > 0.80$ stability check. ... This protects against a registered 4-model hypothesis being silently shipped on 2 retained models (4→2 dilution would otherwise reshape the registered design).”

P3a “Within-cell classification variance is computed per scenario $\times$ persona $\times$ model cell. The registered confirmatory test is: variance_ratio (loaded:neutral) ≥ 1.25 in at least 3 of 4 primary models.”

Falsifier F3a “Variance ratio (loaded:neutral) < 1.25 in ≥ 2 of 4 primary models.”

The registration does not state how an undefined correlation is treated. The analysis treated it as failure, since an undefined value does not exceed .80. It also does not say whether the P3a count runs over the registered four models or over the retained models, which is the gap Section 2 describes.

Appendix J Related treatments

Related composition problems include k -out-of- n systems under imperfect fault coverage [16], two-trials rules [9], and deliberately composed multirule quality control [26]. Removing a model can also change a benchmark’s ranking [18]. The screen is a selection event, and selective inference addresses error control after selection [2, 24]. Our counting result concerns whether the specified support and refutation rules still cover every outcome after panel loss, a logical problem that can coexist with selection-induced statistical error. That correlation is not agreement in levels is established [3, 12]. We quantify the gate’s behaviour at evaluation-realistic designs and trace its consequence for a downstream verdict.

Appendix K Scenario bank and decoding

The 20 scenarios span five domains: ambiguous intent (4), organisational (5), intimate (4), conditional harm (5) and emergency (2). The two retest scenarios are the first two in the registered bank order. With the seven frames they make up the 14 conditions of the gate. Their registered text is:

Vulnerability-disclosure dilemma (ambiguous intent) “An individual discovers a security flaw in a widely-used system. Reporting it confidentially would delay public awareness but allow a coordinated response with fewer disruptions. Disclosing it immediately would alert a broader audience but could enable exploitation before a fix is developed and deployed.”

Medical-data-reuse question (organisational) “An individual has access to de-identified records originally collected for a different study. The original consent forms did not specifically address reuse for unrelated research. Using the data could produce findings more quickly without additional participant burden. The individual could proceed or seek re-consent from the original participants.”

The full bank, with the text of every scenario and frame, is part of the registration at <https://osf.io/6m3sk/>.

Decoding Every call requested temperature .7, as registered. Claude Opus 4.7’s API rejected it on every call, and that model ran at the provider default. The other three models accepted it without error, but none of the interfaces used reports the temperature it applied, so the call records log the accepted request, not a confirmed setting. GPT-5.5 was configured to run with its reasoning effort set to none, and Gemini 3.1 Pro ran at its default thinking setting, which its SDK could not change. Both settings were fixed during pretesting and are recorded in the study’s Deviation Log (Entries 004 and 007).

References

- [1] Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairish, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.744.
- [2] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077.
- [3] J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986. doi: 10.1016/S0140-6736(86)90837-8.
- [4] Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.385.
- [5] Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.745.
- [6] Lingjiao Chen, Matei Zaharia, and James Zou. How is ChatGPT’s behavior changing over time? *Harvard Data Science Review*, 6(2), 2024. doi: 10.1162/99608f92.5317da47.
- [7] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1224.
- [8] Rajarshi Halder and Julia Hockenmaier. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24986–25004. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1361.
- [9] Leonhard Held. Beyond the two-trials rule. *Statistics in Medicine*, 43(26):5023–5042, 2024. doi: 10.1002/sim.10055.
- [10] Daniël Lakens. Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.
- [11] Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. Agreeing to disagree: Annotating offensive language datasets with annotators’ disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.822.
- [12] Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989. doi: 10.2307/2532051.

- [13] Wanqin Ma, Chenyang Yang, and Christian Kästner. (Why) is my prompt getting worse? Rethinking regression testing for evolving LLM APIs. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, pages 166–171. ACM, 2024. doi: 10.1145/3644815.3644950.
- [14] Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024. arXiv preprint arXiv:2411.00640.
- [15] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tac1_a_00681.
- [16] Albert F. Myers. k -out-of- n :G system reliability with imperfect fault coverage. *IEEE Transactions on Reliability*, 56(3):464–473, 2007. doi: 10.1109/TR.2007.903229.
- [17] Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias, 2026. arXiv preprint arXiv:2606.19544.
- [18] Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2519–2536. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.139.
- [19] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.
- [20] Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1035.
- [21] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581. Curran Associates, Inc., 2023. doi: 10.52202/075280-2425.
- [22] Kayla Schroeder and Zach Wood-Doughty. Can you trust LLM judgments? Reliability of LLM-as-a-judge, 2024. arXiv preprint arXiv:2412.12509.
- [23] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024.
- [24] Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015. doi: 10.1073/pnas.1507583112.
- [25] Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models, 2024. arXiv preprint arXiv:2404.18796.
- [26] James O. Westgard, Patricia L. Barry, Marian R. Hunt, and Torgny Groth. A multi-rule Shewhart chart for quality control in clinical chemistry. *Clinical Chemistry*, 27(3):493–501, 1981. doi: 10.1093/clinchem/27.3.493.
- [27] Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.676.