RULES MOVE VERDICTS MORE THAN PARAPHRASES: CONTROLLED EVALUATION OF LLM-JUDGE PANELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agreement does not reveal whether language-model judges follow a stated rule, express an interpretation, or have evidence about an answer. We separate these questions through controlled comparisons. On 160 authored cases and four judges, reversing an explicit rule changes $462/469$ complete ambiguous judge-case verdicts; paraphrasing the same rule changes $4/469$. The assigned-profile contrast is 90.04 percentage points, with completion bounds $[83.59, 97.46]$. A prospectively fixed extension reuses all cases with concurrent explicit controls. With neither rule supplied, an absence notice raises undetermined responses by 33.59 points over filler and turns the 41 binary-unanimous filler panels, among 128 ambiguous cases, into mixed responses. This does not make every compatible binary verdict wrong. A 96-case wording follow-up finds positive aggregate undetermined differences for every notice against both filler and convention mention. However, convention mention also eliminates binary unanimity, notice magnitudes vary, and one wording induces undetermined responses on $23/128$ determinate controls. Exploratory vote analysis shows that broken unanimity need not remove a binary majority. A separate matched-call comparison retains its observed accuracy tie, not equivalence; older elicitation and checking studies retain their registered nulls, uncertainty and harms. These findings concern fixed cases and judge rosters. Replay code and data are available in the anonymous repository.

1 INTRODUCTION

Language models increasingly evaluate other models' answers. Multiple judges can reduce some individual biases (Verga et al., 2024), yet share errors (Goel et al., 2025; Kim et al., 2025; Kohli, 2026). Exact agreement need not establish validity (Norman et al., 2026); agreement with human preferences can also poorly indicate factual and logical correctness (Tan et al., 2025). An evaluator still needs to distinguish following a stated decision rule, expressing an interpretation, and obtaining evidence about the candidate.

One case, three policies. Judge preferences can favor style (Feuer et al., 2025), and task models—not judge panels—show prompt-format sensitivity (Sclar et al., 2024). Our comparison separates wording from rule content. One task, input and candidate receive three length-matched policy notes: two state the same rule in different words; the third states the opposite rule, changing the executable reference label (Figure 1). Across 160 cases and four judges, opposite rules change $462/469$ complete ambiguous triplets, while same-rule paraphrases change $4/469$. The assigned-profile contrast is 90.04 points, bounded by $[83.59, 97.46]$. This identifies local responsiveness to the stated rule, not population coverage or a hidden interpretation.

No policy is not another paraphrase. Recognizing missing information is distinct from answering: models may identify ambiguity when asked without requesting clarification (Su & Cardie, 2026), and underspecified legal queries test whether material omissions prevent a premature conclusion (Vincent et al., 2026). Reusing our cases with concurrent explicit controls, an absence notice raises undetermined replies by 33.59 points over filler and changes all 41 filler binary-unanimous panels to mixed responses. A further wording study supports positive aggregate notice effects but limits the interpretation: merely mentioning conventions also removes binary unanimity, and one notice over-abstains on determinate controls.

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.

One case, three policy slots. The task, the input, the candidate and the instruction header are identical in every arm; only the twelve-word policy slot changes. Two slots state the same rule and share an executable reference; the third states the opposite rule, which moves that reference.

Policy slot	Twelve words shown to the judge	Reference for scoring
Stated	Use JSON null as the average whenever no nonnegative group values qualify.	`CORRECT`
Reworded	An empty qualifying subset receives null for its reported average result value.	`CORRECT`
Opposite	Use numeric zero as the average whenever no nonnegative group values qualify.	`INCORRECT`

Task: for each group, return the average of its nonnegative values, preserving group order. Input {`"groups":[[-74,-76],[78,-2,80]]`} with candidate `[null,79.0]`. All four judges returned `CORRECT` under both same-rule slots and `INCORRECT` under the opposite one.

Verdict changes across the cohort, over judge-case triplets with three replies:



Figure 1: **The stated rule moves verdicts far more than rewording.** The case is selected for illustration, not representativeness; the policy slot is a model input, while the reference label is used for scoring and is never shown to the judge. Alignment spaces are omitted here. The bars are the measurement: on the 469 judge-case triplets that returned all three verdicts, reversing the rule changes 462 verdicts and rewording it changes 4. Over all 512 assigned judge-case profiles, which keep every missing or invalid reply as unknown rather than imputing a vote, the observed contrast is 90.04 percentage points, with exact allowed-label-completion bounds $[83.59, 97.46]$. These are judge-case comparisons, not all-four-panel outcomes. The illustrated mechanism is the strongest of the eight authored mechanisms, whose lower bounds are all positive and as low as 65.62 points (Table 13).

We make three contributions, rather than proposing a new voting rule. First, the explicit-rule test isolates rule reversal from paraphrase. Second, the omission and wording controls separate undetermined responses from unanimity, majority decisions and reference-relative correctness. Third, the decision/expression/added-evidence storyline connects these findings to exploratory vote composition, a matched-call accuracy tie, and retained registered nulls and check tradeoffs. Table 1 distinguishes their cohorts and status. An accompanying anonymous supplemental ZIP replays the controlled studies and exploratory votes from saved outputs; it does not regenerate provider responses (Appendix H).

2 Related work

Bias and agreement. Position and self-enhancement bias (Zheng et al., 2023), self-preference (Panickssery et al., 2024), inconsistent ratings (Stureborg et al., 2024) and variable agreement with humans (Bavaresco et al., 2025) are distinct problems. Our reference is the executable consequence of a declared rule, not preference agreement alone.

Dependence and aggregation. Self-consistency aggregates reasoning paths from one model (Wang et al., 2023), unlike a multi-model jury. CAPA measures shared mistakes (Goel et al., 2025); cross-provider analyses document dependence (Kim et al., 2025), and effective-vote analyses connect panel entropy with human-label ambiguity (Kohli, 2026). Marginal gains from sentiment majorities challenge independence assumptions (Lefort et al., 2024); shared confounders can make aggregation amplify error (Zhao et al., 2026). Our vote analysis describes decision consequences, not another independence estimate or aggregation method.

Perturbations and references. Fixed-content rhetorical rewrites can change AI-reviewer scores (Li et al., 2026b). LLMBar instead tests instruction-following discrimination (Zeng et al., 2024), and RevisEval constructs response-adapted references (Zhang et al., 2025). Representation-as-a-Judge studies representation probes as an alternative to prompted judging (Li et al., 2026c); it is not

evidence for a fixed-content wording effect here. We hold the candidate fixed while changing the rule and therefore its reference label, with a same-rule paraphrase as control.

Ambiguity and expression. Finding ambiguity differs from ordinary annotation (Pradhan et al., 2022), and disagreement can be resolvable or irresolvable (Schaekermann et al., 2018). Rating criteria may admit multiple reasonable answers (Guerdan et al., 2025). These distinctions motivate our separate measurements of stated verdicts and elicited interpretations; neither identifies a judge's prior internal state.

Selective and calibrated evaluation. Reject-option and selective prediction formalize withholding decisions (Chow, 1970; Geifman & El-Yaniv, 2017); escalation is a further decision (Jung et al., 2025). Calibrated judge thresholds can route uncertain cases to retrieval (Badshah et al., 2026). Conformal risk control bounds expected monotone loss using calibration (Angelopoulos et al., 2024), while noisy-judge testing calibrates judge error before certification (Feng et al., 2026). Finite-budget panel work compares low-dimensional and joint-table calibrators, accounting for labels spent on path construction, fitting and validation (Zhu et al., 2026). If a judge is no more accurate than the evaluated model, debiasing cannot more than halve the required ground-truth sample (Dorner et al., 2025). Our local completion bounds neither implement these procedures nor replace their assumptions.

Generator–evaluator relatedness. Preference leakage concerns relatedness between a synthetic-data generator and the later evaluator—same model, inheritance or model family—not merely exposure to test items (Li et al., 2026a). We do not measure that relatedness.

3 TASKS, PANELS, AND DECISION RULES

The studies ask what a verdict follows, what a reply expresses, and what a further check supports. Table 1 shows which comparison answers each question; the rows are not successive stages of one experiment. A *decision* is the result of applying a stated rule to the replies it accepts. An *expressed interpretation* is read from a rationale or follow-up answer using its named coding instrument. *Added evidence*, such as a supplied condition or an execution run, concerns the candidate within that check's reference and tested setting.

Cases and executable references. Controlled behavioral cases can isolate capabilities hidden by aggregate accuracy (Ribeiro et al., 2020); contrast sets use meaningful edits that typically change the gold label (Gardner et al., 2020). Our edited object is the policy slot. As in verifiable instruction checks (Zhou et al., 2023), correctness has an executable criterion, but here the stated policy determines which reference applies. Ambiguous cases have opposite candidate labels under the two conventions; determinate controls have the same label under both. The fixed roster is GPT-6 Astra, GPT-5.6, Opus 5 and Opus 5.5, from two providers. Eight authored mechanisms supply 160 policy-slot cases (128 ambiguous, 32 determinate) and 96 disjoint matched-call cases. Fresh values and policy wording do not constitute new mechanism families.

A short unit key. An *assigned profile* is one judge and one case: the primary 512 ambiguous profiles are 128 cases times four judges. A *complete triplet* has valid verdicts in all three policy arms; four-judge unanimity requires four valid equal labels. Missing or invalid replies remain unknown, not agreement, correctness or UNDETERMINED (U). Completion bounds allow every unknown reply any permitted label while holding observed labels fixed. They are not sampling confidence intervals.

Panel decisions use different denominators. The older registered paired study conditions on unanimous panels with at least four valid votes from eleven judges. A separate post-hoc diagnostic requires all eleven draw-zero replies and retains fixed assigned denominators. An incomplete or split panel makes no automatic decision, not a correct one. Full rules, rosters, parsers and weightings remain in Appendices A and F.

Table 1: Cohorts, comparisons and status. Prospective means the design was fixed before its calls, not that reused cases were unseen. The policy-slot and matched-call cohorts are disjoint; omission and wording reuse policy-slot cases. The rows are not independent replications or stages of one experiment.

Study	Comparison and unit	Section
Explicit-rule test	Prospective; 160 cases, four judges: rule reversal versus paraphrase in judge–case triplets.	4
Withheld-rule extension	Outcome-informed reuse of all 160 cases; prospectively fixed five-arm acquisition and analysis, with concurrent explicit controls.	4
Vote composition	Post-review exploratory reanalysis: all 128 ambiguous cases and the 41 F-unanimous subset; no new calls.	4
Notice wording	Prospective; 96 balanced reused cases selected without outcome filtering; four notices, filler and convention mention.	4
Matched-call check	Prospective; 96 disjoint cases, four judges: same menu and two calls per procedure, measured in judge–case pairs.	4
Registered paired study	720 items, eleven judges: conditional unanimity and separate post-hoc complete-panel diagnostics.	4
FOLIO convention test	263 paired blocks, eleven judges: registered unanimity and post-hoc decision rules.	4
Disclosure studies	Separate SQL/regex cohorts: model and author codes, menu choices and follow-up answers, with instrument-specific units.	5
Condition, execution, guard and routing	Distinct comparisons of false accusations, missed bugs, flags and routing loss; each keeps its own reference and denominator.	6

4 What changes when the requirement is stated?

We first separate rule responsiveness from paraphrase, then distinguish the effects of omission and wording on individual replies and panel decisions. The retained registered studies and matched-call check answer different questions.

The rule, not its wording. The prospective policy-slot test assigns 1,920 calls. Within a case, task, input, candidate and header are fixed. Specified (S) and paraphrase (P) state the same rule; opposite (A) reverses it and moves the executable reference label. Slots have twelve words and matched ASCII width. On 469 complete ambiguous triplets, S/A changes 462 verdicts (98.5%) and S/P changes 4 (0.9%). Across all 512 assigned ambiguous profiles, the observed contrast is 90.04 points with completion bounds $[83.59, 97.46]$. Every mechanism has a positive lower bound, the smallest 65.62 points.

Checks on the controlled comparison. Determinate verdicts stay unchanged in $119/120$ valid S/A pairs and $118/118$ S/P pairs. Opposite slots differ by only one to eight word edits versus eleven or twelve for paraphrases. The stricter parser, frozen before collection, retains a 73.44-point lower bound on the same responses. All four judges match all three references on $78/128$ *assigned* ambiguous cases; 85 rosters are complete. Neither that completeness count nor a judge-case rate replaces the assigned panel denominator. Joint correctness, acquisition details and full grids are in Appendix B.

Following a reference, not merely changing a label. Verdict changes alone would not establish rule following. Judges match both S/A references on $466/512$ assigned profiles and all S/P/A references on $462/512$. The task and candidate stay fixed, but the reference *label* does not: the opposite policy changes which executable result is required. Policy text is a model input; executable outputs and oracle verdict labels are not. These counts establish reference-relative responsiveness within the authored two-convention family, not correctness under every plausible reading of a natural-language task.

Omission with concurrent explicit controls. The five-arm extension reuses all 160 policy-slot cases. Reuse follows the earlier results, but arms and analysis were fixed before 3200 new calls; no case was filtered by its earlier outcome. S/A reacquire explicit policies under the same header as generic filler F, absence notice N and exact repeat R. N says "This policy note states no controlling convention for the present evaluation task." Twelve-word notes have matched ASCII width. No oracle label or policy menu is supplied, and abstention is allowed, not prescribed. Rounding and sorting require one fixed tie convention, not any tied alternative (Appendix G).

The primary N/S comparison changes $429/496$ valid ambiguous pairs versus $2/122$ determinate pairs. On assigned profiles, this is an 82.23-point ambiguity-specific contrast with completion bounds $[77.54, 85.35]$. Against F, N raises undetermined replies from $203/509$ to $375/511$ valid ambiguous calls: an assigned increase of 33.59 points, $[33.01, 33.79]$, with positive lower bounds in every mechanism (Appendix Table 30).

The panel change is direct. F yields four equal binary verdicts on $41/128$ ambiguous cases across four mechanisms; N yields none, and all 41 become complete mixed panels. The same 124 jointly complete cases give 41 versus zero. Assigned-case bounds on the decrease are $[31.25, 34.38]$ points; R also yields zero binary-unanimous panels. N instead gives 45 all-undetermined, 82 mixed and one incomplete panel. Its 136 remaining binary replies split $68/68$ between the two reference labels; they are not universally wrong or evidence of an internal convention.

This original notice gives zero U on determinate controls and $124/124$ valid reference-correct judgments, or $124/128$ assigned. Both explicit poles remain reference-correct on $484/512$ assigned ambiguous profiles ($484/487$ valid pairs). The wording follow-up below tests how far this specificity extends.

The N/S and N/F contrasts answer different questions. N/S counts any verdict change after removing a supplied rule; N/F changes only the neutral note while leaving both rules absent. Neither turns every remaining binary judgment into an error. The smallest mechanism-level N-minus-F U lower bound is 3.125 points, on rounding ties, so a positive direction does not imply equal magnitudes. N/R differs on $31/507$ valid pairs: an exact-prompt repeat readout, not a new mechanism replication or a policy discovery test.

Exploratory votes: broken unanimity is not a lost majority. The post-review reanalysis of saved F/N/R replies makes no new calls and changes no registered endpoint. Of the 41 formerly unanimous F panels, 15 become one U plus three binary votes under N, 16 become two U plus two binary, and 10 become three U plus one binary. There are no binary-label flips: 15 panels retain a strict majority for their original F label. On all 128 assigned cases, strict binary majorities number 54 (F), 16 (N) and 14 (R), with 125, 127 and 124 complete rosters. A strict majority here requires three or four equal votes in a complete four-judge roster. Two-U/two-binary panels make no decision under that rule; they are not counted as correct decisions. Dropping Astra, GPT-5.6, Opus 5 or Opus 5.5 gives three-judge binary-unanimity counts F→N of $43 \to 6$, $43 \to 0$, $44 \to 0$ and $48 \to 10$, respectively, each on all 128 assigned frames. Thus the unanimity result depends on the roster and decision rule. These are decisions, not universal errors or deployment gains (Appendix G.4).

Wording controls change the interpretation. A prospective follow-up uses a balanced 96-case subset of the same cases (64 ambiguous, all 32 determinate), selected by frozen metadata-only ranking without outcome filtering. It reacquires F, original notice N_0, three paraphrases N_1–N_3, and convention mention M, which neither supplies a rule nor declares its absence. All notes have twelve words and 99 ASCII characters under the same output contract. F and N_0 reproduce the selected cases' earlier F and N prompts byte for byte, but all six arms are newly acquired in shuffled case–judge blocks. No earlier response is substituted as a concurrent control. Because the follow-up retains the prior cases and reference family, it tests wording on this benchmark rather than generalization to new task mechanisms. There are $2280/2304$ valid assignments over 2310 attempts, with 23 parser failures and 1 acquisition failure. Table 2 keeps valid and assigned denominators visible.

Each notice has a positive aggregate U difference versus both F and M, including every adverse-completion lower bound (all eight contrasts in Table 3). Assigned differences range 34.38–59.77 points against F and 23.05–48.44 against M; the smallest completion lower bound across these eight comparisons is 21.88 points. But the assigned notice U rates range from 73.44% to 98.83%: the magnitude is not wording-invariant.

Table 2: Wording controls qualify the omission finding. All notices and convention mention M have zero binary-unanimous panels, unlike filler F; N_2 also incurs substantial determinate abstention. U is UNDETERMINED; control "correct" means reference-correct. Slashed counts explicitly retain valid and assigned frames. Binary unanimity requires four valid equal binary labels on one of the 64 assigned ambiguous cases.

Arm	Ambiguous replies U / valid / assigned	Determinate controls U / correct / valid / assigned	Binary unanimity observed / assigned
F	$100/252/256$	$0/124/125/128$	$19/64$
M	$129/253/256$	$1/125/126/128$	$0/64$
N_0	$188/254/256$	$0/126/126/128$	$0/64$
N_1	$222/255/256$	$1/126/127/128$	$0/64$
N_2	$253/256/256$	$23/101/124/128$	$0/64$
N_3	$200/256/256$	$5/121/126/128$	$0/64$

Table 3: Every notice raises aggregate U relative to both controls. All eight comparisons use 256 assigned ambiguous judge–case profiles per arm. Completion bounds retain observed labels and allow every nonvalid reply any permitted label; they are not confidence intervals. Positive aggregate lower bounds do not establish equal wording effects or model-uniform separation.

Contrast	Assigned difference (pp)	Completion bounds (pp)
N_0 versus F	34.38	$[32.81, 35.16]$
N_0 versus M	23.05	$[21.88, 23.83]$
N_1 versus F	47.66	$[46.09, 48.05]$
N_1 versus M	36.33	$[35.16, 36.72]$
N_2 versus F	59.77	$[58.20, 59.77]$
N_2 versus M	48.44	$[47.27, 48.44]$
N_3 versus F	39.06	$[37.50, 39.06]$
N_3 versus M	27.73	$[26.56, 27.73]$

What the controls rule out. M and *every* notice have zero binary-unanimous panels out of 64, with completion-count bounds $[0, 0]$, versus F's $19/64$, bounded by $[19, 20]$. Loss of unanimity is therefore *not specific* to explicitly declaring absence. The largest U rate, N_2, also gives $23/128$ determinate U and only $101/128$ assigned reference-correct judgments. F, M, N_0, N_1 and N_3 give respectively $0, 1, 0, 1, 5$ control U. Nor does every model/wording separate: Opus 5.5 gives $61/64$ ambiguous U for both N_3 and M. These controls rule out a best-wording or uniform-effect claim, not the positive aggregate U contrasts.

Retained registered panel evidence. The older eleven-judge study remains a separate comparison. Its registered conditional endpoint is $0/146$ versus $28/106$ on 214 blocks ($+26.4$ points; $p = 1/20{,}001$). Constraint checking contributes 22 disagreements, SQL six; arithmetic contributes none. The author-class exclusion removes all SQL pairs. Unlike the new matched slots, the constraint intervention rewrites the header and all 30 definitions, with no length-matched placebo.

The intended `--k 1` collection sampled a hash-selected third five times; the frozen selector retains the last valid reply. Requiring all eleven draw-zero replies gives $0/256$ versus $7/256$ on 214 blocks. Six events come from one SQL class, and grouping gives $p = 0.25$. Constraint completeness is $4/100$ withheld versus $36/100$ stated. These are not independent confirmations of the new omission contrast. Appendix A.3 preserves the full original description, retry results and references; Table 9 keeps the registered and complete-panel endpoints distinct.

FOLIO retains its registered null. On public formal premises (Han et al., 2024), the all-eleven unanimity endpoint remains $0/263$ in both arms. The later six/seven-vote policies are post-hoc, convention-relative analyses, not a replacement for that null: without a convention, `FALSE` can mean "not entailed" rather than "refuted." All 54 comparisons, coverage values and interpretation limits remain in Appendix E.2 and Table 28.

Same menu, same planned calls, observed accuracy tie. On 96 disjoint cases, both procedures receive the same exhaustive two-policy menu and spend two calls per judge-case. The baseline jointly assesses both policies twice; separate-policy assessment selects one per call. The same symmetric rule keeps agreeing valid labels, returns U for differing valid labels, and marks any nonvalid pair nonvalid, not abstaining. Both attain 94.34% class-balanced accuracy: 0.00 points difference, stability interval $[-0.98, 1.17]$, completion bounds $[-5.47, 4.88]$. This is neither superiority nor equivalence, nor a tie obtained by deferring everything. Both procedures reach the case target on $241/256$ ambiguous pairs and $121/128$ determinate pairs. The target is U when the supplied menu leaves opposite candidate labels and the common binary label when both policies agree.

Separate-policy calls give two reference-aligned conditional judgments on $237/256$ ambiguous pairs. Another 4 correct aggregate deferrals contain a conditional abstention: a correct aggregate target does not by itself show that both conditional judgments were reference-aligned. That makes a conditional judgment scorable, not uniquely informative: baseline prose was not coded for that content, and downstream benefit was not measured. Planned calls and caps, not realized tokens, are matched; no cost saving is claimed. Full rates, near-deterministic baseline repeat results and mechanism tables remain in Appendix B.

5 WHAT DO THE JUDGES SAY ABOUT AN OPEN REQUIREMENT?

We next ask whether judges describe a missing requirement, and whether asking about it changes their response. These experiments use separate SQL and regular-expression cohorts, not the studies above. Rationale codes, author codes and parsed follow-up answers do not measure one quantity (Appendix C).

Rationales, questions, and human codes. The rubric counts a rationale that names a missing requirement, states an assumption, or asks for clarification. The baseline requested one short main reason; its parser retained the first rationale line, capped at 400 characters. Absence in that format is not inability to describe another reading. In the eleven-judge menu study, panel silence is 72.2% under the plain prompt, 70.8% when an UNDERSPECIFIED option is offered, and 51.4% when judges are invited to flag an open requirement. The last point estimate misses the registered cutoff below 50%; its $[37.5, 65.3]\%$ interval spans that cutoff. The larger silence reduction uses PERMITTED; the accuracy comparison uses OFFERED. They are not a joint accuracy-and-disclosure gain.

The author check changes the estimated rates. Two authors coded 200 sampled rationales. Nominal $\alpha = 0.715$ passes the registered 0.667 gate on this positive-enriched sample. The registered union estimates 16.6% ambiguous-arm disclosure on the source frame, versus 6.5% under model codes; intersection gives 5.7%. These differ from later menu re-coding. Positive-class agreement remains limited: frame-weighted human–human Jaccard is 0.239, union–model Jaccard 0.251. The ambiguous-minus-placebo interval $[-11.7, 17.6]$ includes zero. The frozen outcome is SUPPORTS-REVISED. This scattered sample does not validate complete-panel silence or follow-up arms; individual-coder estimates and sampling details remain in Appendix C.1.

A question after the verdict changes the answer. In the same-conversation SQL probe, $275/780$ under-specified cells move from no coded flag to a parsed `OPEN` answer, against one reverse; all 36 base questions move the same way. Explicit controls give 8.2% `OPEN`. The follow-up introduces the choice and changes the measurement instrument: it establishes elicitation, not prior awareness. The preregistered two-turn regular-expression probe gives a control-corrected $+41.6$ points $[+35.2, +47.8]$, and scoring all 186 missing cells adversely still leaves $+24.9$ points. A separate one-turn DIRECT-minus-OFFERED study has a $+49.9$-point raw gap but only $+17.9$ after control correction, below its registered 20-point threshold, and its "explicit" controls retain a secondary ambiguity identified in 142 rationales (Appendix C.6).

An elicited statement makes an interpretation explicit when asked for; it neither establishes that the judge held it earlier nor verifies the candidate.

6 WHAT DOES AN ADDITIONAL CHECK ESTABLISH?

Verifier-guided selection uses an additional judgment to choose candidates (Kwok et al., 2026). Our added checks have narrower, separately measured effects. Table 4 pairs each benefit with its cost, failed check or missing control; rows are not one intervention pipeline.

Table 4: Added checks can help one outcome and hurt another. Each row pairs a measured benefit with its cost, missing control or failed check. No row is a validated policy, a missing control is not an observed harm, and rows are not additive across cohorts. Loss per 100 is a retrospective routing loss under a stated cost ratio, not a conditional error rate.

Intervention	Measured benefit	Measured cost or missing control
Supply the missing task condition (five arms, 28 variants)	Per-judge false alarms $0.622 \rightarrow 0.107$; panel error $0.804 \rightarrow 0$.	Separate control arms worsen false alarms ($+0.095$ placebo, $+0.122$ asserting the edge case present); these are not harms of the helpful condition.
State the invariant on a screened cohort ($155/160$ retained)	Tie-scored panel errors on 80 reference-correct cases fall $23.5 \rightarrow 0$; neither arm misses any of the 75 buggy cases.	The five exclusions are label changes, not observed misses; the gold-free preflight failed, so safe uses cannot be identified automatically.
Execute the candidate (138 of 224 variants)	Per-critic error $15.8\% \rightarrow 7.1\%$.	Buggy misses rise $3.0\% \rightarrow 12.8\%$; tests filter on the exact clean reference, so the clean-side zero is a same-reference ceiling.
Adjudicate with an execution guard (642 implementations, 112 mutations)	Flags $22.1\% \rightarrow 6.4\%$ (-15.73 pp, $[-22.10, -9.51]$) at 93.3% decision coverage.	Recovery $92.9\% \rightarrow 89.3\%$: four catches lost. Visible-test passing does not certify correctness; no matched bare-vote baseline exists.
Route on an elicited flag (144 SQL items, $k = 10$)	Flags $11/11$ wrongly rejected items; beats vote margin at matched coverage.	Catches only $3/9$ missed bugs; loss per 100 is 53.5 (gate) and 60.4 (gate plus rescue) against 50.0 for deferring everything, before inference cost. The registered matched-coverage confidence comparison fails.

Supplying a task condition. A preregistered five-arm SQL test finds fewer false alarms when an unstated edge case is ruled out. Placebo and wrong-way condition arms worsen them; these are separate controls, not harms of the helpful condition. The broader invariant study retains $155/160$ items after five wrong-join candidates become reference-equivalent. The five exclusions are not observed new misses. Its gold-free preflight failed, so the screened gain does not identify safe uses automatically. The original stratum, label screen and control outcomes are retained in Appendix D.2.

Execution changes both error channels. Execution reduces per-critic error but increases buggy misses on the 138 variants complete in both arms. Tests filter on the exact clean reference, making the clean-side zero a same-reference ceiling. A separate guard reduces flags but four catches are lost on certified mutations. Visible-test passing does not certify correctness: removed flags are not all known false accusations, and no matched bare-vote baseline was collected. Appendix D.4 preserves both error channels and coverage.

A disclosure is not automatically a useful escalation signal. The elicited SQL flag beats vote margin at matched coverage but fails its registered comparison with mean stated confidence. Its error-channel split favors wrongly rejected items over missed bugs. Both later flag-routing policies lose to defer-all at the primary cost ratio before inference cost. A separate retrospective confidence comparison gives -25.69 points with nested interval $[-50.69, 7.66]$, crossing zero; it is not the registered matched-coverage comparison (Appendix D.3). Together with the matched-call tie in Section 4, these results support specific conditional checks, not a general escalation policy.

7 What can an evaluator conclude?

Decisions: identify the rule and denominator. Section 4 distinguishes rule responsiveness from omission: broken unanimity can retain a binary majority, and convention mention also disrupts it. Report assigned coverage and nondecisions alongside conditional error; compatible binary labels are not universally wrong.

Expression: identify the instrument. Prompted interpretations do not establish prior awareness, candidate correctness or independently validated panel silence; author-coding uncertainty remains (Section 5).

Added evidence: identify what the check establishes. Table 4 pairs benefits with harms and missing comparators. Conditions define tasks, selected policies make judgments conditional, and execution tests behavior within its reference and tested setting; none certifies arbitrary candidates.

8 Conclusion

Our controlled studies show how task specifications shape the decisions of LLM-judge panels. With an explicit rule, verdicts respond strongly to changes in that rule and remain largely stable under matched paraphrases. When the rule is omitted, convention cues change expressed uncertainty and reshape the votes that feed a panel's decision. Examining those votes under unanimity and majority rules connects individual responses to their consequences for the panel.

Together, these findings offer a practical way to read model agreement. The task specification defines the criterion being judged, the decision rule determines how responses are combined, and follow-up questions or checks contribute further information. Evaluating these elements together gives a clearer account of what a panel decided and the evidence supporting that decision.

Use of AI statement

Generative AI assisted synthetic-data construction, method implementation, data cleaning/reformatting, qualitative coding, and simulated review. It also assisted literature search/synthesis, artifact creation, and manuscript editing. AI-assisted code and results were checked through tests, executable references and offline replay, and citation claims against sources. The authors take responsibility for the final text, claims, code and data.

Reproducibility statement

Methods and appendices detail the checks. The anonymous replay repository reproduces $8{,}960$ saved assignments without model calls (Appendix H).

REFERENCES

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, pp. 55198–55218, 2024.

Sher Badshah, Ali Emami, and Hassan Sajjad. Judge, retrieve, or abstain: Uncertainty-guarded LLM judging with provable risk guarantees. *arXiv preprint arXiv:2608.17994*, 2026. URL https://arxiv.org/abs/2608.17994.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, 2025. doi: 10.18653/v1/2025.acl-short.20.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *International Conference on Learning Representations (ICLR)*, 2025. Oral; arXiv:2410.13341.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In *International Conference on Learning Representations (ICLR)*, pp. 4643–4691, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/08a67eb74bfb9ded9e949dd52973e997-Abstract-Conference.html. arXiv:2601.20913.

Benjamin Feuer, Micah Goldblum, Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, and John P. Dickerson. Style outweighs substance: Failure modes of LLM judges in alignment benchmarking. In *International Conference on Learning Representations (ICLR)*, pp. 11552–11582, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1eb36d07ebb13be16ddbda679a95018b-Abstract-Conference.html. arXiv:2409.15268.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19621–19678. PMLR, 2025. URL https://proceedings.mlr.press/v267/goel25b.html. arXiv:2502.04313.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova. Validating LLM-as-a-judge systems under rating indeterminacy. In *Advances in Neural Information Processing Systems*, volume 38, pp. 124475–124543. Neural Information Processing Systems Foundation, Inc., 2025. doi: 10.52202/085713-3751. URL https://doi.org/10.52202/085713-3751. arXiv:2503.05965.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander R. Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1229. URL https://aclanthology.org/2024.emnlp-main.1229/.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations (ICLR)*, pp. 3101–3125, 2025. arXiv:2407.18370.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30038–30066. PMLR, 2025. URL https://proceedings.mlr.press/v267/kim25e.html.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. *arXiv preprint arXiv:2605.29800*, 2026. URL https://arxiv.org/abs/2605.29800.

Jacky Kwok, Shulu Li, Pranav Atreya, Yuejiang Liu, Yixing Jiang, Chelsea Finn, Marco Pavone, Ion Stoica, and Azalia Mirhoseini. LLM-as-a-verifier: A general-purpose verification framework. *arXiv preprint arXiv:2607.05391*, 2026. doi: 10.48550/arXiv.2607.05391. URL https://arxiv.org/abs/2607.05391.

Baptiste Lefort, Eric Benhamou, Jean-Jacques Ohana, Beatrice Guez, David Saltiel, and Thomas Jacquot. Examining independence in ensemble sentiment analysis: A study on the limits of large language models using the Condorcet jury theorem. *arXiv preprint arXiv:2409.00094*, 2024.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: A contamination problem in LLM-as-a-judge. In *International Conference on Learning Representations (ICLR)*, pp. 60486–60511, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6297baf3f5c98146b62dd3a1bffe068e-Abstract-Conference.html. arXiv:2502.01534.

Ming Li, Chenguang Wang, Xirui Li, Xinyue Zeng, Dianqi Li, Peng Shi, Dawei Zhou, and Tianyi Zhou. How can rhetoric reward-hack AI reviewers? dissecting rhetorical sensitivity in AI-based peer review. *arXiv preprint arXiv:2608.08975*, 2026b.

Zhuochun Li, Yong Zhang, Ming Li, Yuelyu Ji, Yiming Zeng, Ning Cheng, Yun Zhu, Yanmeng Wang, Shaojun Wang, Jing Xiao, and Daqing He. Rethinking LLM-as-a-judge: Representation-as-a-judge with small language models via semantic capacity asymmetry. In *International Conference on Learning Representations (ICLR)*, pp. 77711–77737, 2026c. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7d9173a6010bed38c359fa012becd4da-Abstract-Conference.html. arXiv:2601.22588.

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. *arXiv preprint arXiv:2606.19544*, 2026.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802, 2024. doi: 10.52202/079017-2197. URL https://arxiv.org/abs/2404.13076.

Vivek Krishna Pradhan, Mike Schaekermann, and Matthew Lease. In search of ambiguity: A three-stage workflow design to clarify annotation guidelines for crowd workers. *Frontiers in Artificial Intelligence*, 5:828187, 2022. doi: 10.3389/frai.2022.828187. arXiv:2112.02255.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19, 2018. doi: 10.1145/3274423. Article 154.

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 *International Conference on Learning Representations (ICLR)*, pp. 25055–25083, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6c0e99d736da621403018ca7b32b1a4d-Abstract-Conference.html. arXiv:2310.11324.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

Jinyan Su and Claire Cardie. Knowing but not showing: LLMs recognize ambiguity but rarely ask clarifying questions. *arXiv preprint arXiv:2605.25284*, 2026. URL https://arxiv.org/abs/2605.25284.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations (ICLR)*, pp. 63277–63303, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9e720fce64f91114c49cfd640d821da3-Abstract-Conference.html. arXiv:2410.12784.

Pat Verga, Sebastian Hofstatter, 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. *arXiv preprint arXiv:2404.18796*, 2024. URL https://arxiv.org/abs/2404.18796v2.

Samuel J. Vincent, Daniel Calloway, Fangyi Yu, Andrew M. Bean, and Nabeel Seedat. InsufficiencyBench: Evaluating LLM legal advice on underspecified user queries. *arXiv preprint arXiv:2608.20220*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations (ICLR)*, pp. 40193–40219, 2024.

Qiyuan Zhang, Yufei Wang, Tiezheng Yu, Yuxin Jiang, Chuhan Wu, Liangyou Li, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and Chen Ma. RevisEval: Improving LLM-as-a-judge via response-adapted references. In *International Conference on Learning Representations (ICLR)*, pp. 99866–99893, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f7ed2318cec3540ca267c3ede12d82fe-Abstract-Conference.html. arXiv:2410.05193.

Jitian Zhao, Changho Shin, Tzu-Heng Huang, Satya Sai Srinath Namburi GNVV, and Frederic Sala. CARE: Confounder-aware aggregation for reliable LLM evaluation. *arXiv preprint arXiv:2603.00039*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. URL https://arxiv.org/abs/2311.07911.

Bin Zhu, Yi Xie, and Yanghui Rao. A finite-calibration regime map for LLM judge panels. *arXiv preprint arXiv:2606.01034*, 2026. URL https://arxiv.org/abs/2606.01034.

Table 5: The rules and cohorts used in this paper. A rule fixes both numerator and denominator, so two rules applied to the same replies answer different questions, and "no event" never means "correct": an incomplete or split panel makes no automatic decision. The first two rows are the two controlled studies reported here, in judge-level units; the rows below them are separate cohorts with their own rosters and instruments, not later stages of one experiment.

Rule or cohort	What it counts	Denominator, coverage and scope
Policy-slot change (controlled test)	Whether one judge's verdict on one case differs between two policy arms, and whether it matches that arm's executable reference.	512 assigned ambiguous judge-case profiles; 469 returned all three replies. A missing or invalid reply stays unknown and is never imputed, so bounds accompany the observed contrast. No panel vote is taken.
Matched-call policy selection (controlled check)	Whether a procedure's aggregate decision matches the case target: abstain when two policies disagree, decide when they do not.	384 assigned judge-case pairs per procedure (256 ambiguous, 128 determinate). Agreeing valid labels are kept, differing valid labels give UNDETERMINED, and a nonvalid reply makes the pair nonvalid rather than abstaining.
Registered conditional (registered paired study)	Reference disagreement among panels unanimous on their valid votes, at least four of eleven valid, last valid reply per cell.	Eligible unanimous panels only. Arm coverage differs, so the arms can compare different panel sizes.
Complete panel, post-hoc (registered paired study)	The same disagreement when all eleven draw-zero replies are valid and unanimous.	All 256 matched items per arm in 214 blocks; constraint and SQL blocks hold one item per side, arithmetic blocks two.
Fixed retry completion, post-hoc (registered paired study)	The same rule after invalid first votes are filled from existing 2048-token replies.	Same items; collects nothing new and does not isolate a token-budget effect.
Registered unanimity and post-hoc thresholds (FOLIO)	Identical verdicts from all 11 judges; post-hoc, at least k of eleven after every reply is recognized.	Fixed 263 blocks. Panels below k abstain, so decisions and disagreements must be read together.
Disclosure cohorts (SQL, regular expression)	Model-coded rationale flags, author codes on a sampled frame, and parsed answers to a follow-up question.	Separate eleven-judge cohorts; first-turn codes and second-turn answers are different instruments, and a coded absence is absence in that response format.
Intervention cohorts (condition, execution, guard, routing)	Reference-relative false accusations, missed bugs, raised flags, abstentions and routing loss.	Separate screened SQL and six-critic execution cohorts, each with its own reference and denominators; costs are not matched across them.

A STUDY DESIGN, COMPLETENESS, AND THE REGISTERED COMPARISON

What each study measures. Table 5 lists every rule and cohort used in the paper with the denominator and scope limit that travels with it, and Table 6 maps every claim in the body to the cohort, measurement and denominator it rests on, so that a rationale code, a follow-up answer and a panel verdict are not read as three estimates of one quantity.

Design and endpoint. The registered paired study contains 720 items from four constructed families. Constraint checking has 100 strict/vague twins with the same record and executed reference. SQL has 72 explicit/ambiguous pairs and 48 intent-preserving placebo items. Arithmetic judging has 42 paired base clusters (168 items); a second arithmetic family contributes 160 unpaired items to class summaries only. The frozen test flips stated- and withheld-side collections within 214 eligible blocks. Placebo joins the stated side. Its endpoint is disagreement with the construction reference *conditional on unanimity*; the two arms can therefore have different unanimity denominators. This is not an unconditional risk contrast on a fixed accepted set.

Table 6: The studies measure different outcomes. A rationale code, a follow-up answer, and a panel verdict cannot be pooled as estimates of the same quantity.

Study	Measurement	Unit and scope
Policy-slot test	Verdict changes under an opposite rule versus a same-rule rewording.	160 constructed cases, four judges; 512 assigned ambiguous judge–case profiles, 469 complete triplets. All-four panel outcomes are reported separately.
Matched-call check	Case-target accuracy and separately scorable conditional judgments.	96 disjoint cases, four judges; 384 assigned judge–case pairs per procedure, with the same two-policy menu and call budget.
Registered paired study	Reference disagreement conditional on unanimity; separately, complete-panel events.	720 constructed items, eleven judges; selected-panel and assigned-item denominators shown separately.
FOLIO convention test	Registered unanimity; post-hoc cutoff decisions on the same replies.	263 paired neutral blocks, eleven fixed judges; all replies must be recognized before a policy decides.
SQL disclosure	Model-coded rationale flags versus parsed `OPEN` answers after a verdict.	Eleven-judge rationale/probe cohorts, distinct from the earlier twelve-judge paired SQL study.
Author coding check	Registered union of two author codes; individual and intersection sensitivities.	200 sampled rationales from the original coding frame; design-weighted rates, not complete-panel estimates.
Execution and task conditions	Reference-relative false accusations, bug catches, flags, and abstentions.	Separate screened SQL and six-critic execution cohorts; each intervention keeps its own denominator and error channels.

Table 7: Registered paired study under construction labels. Counts are reference-disagreeing unanimous panels over unanimous panels in that class. The paired primary below uses different, paired-side denominators.

Family	Stated	Placebo	Withheld	Paired blocks
Constraint checking	$0/16$	—	$22/41$	100
SQL	$0/44$	$0/30$	$6/33$	72
Arithmetic judging	$0/56$	—	$0/32$	42
Second arithmetic family	$0/8$	$0/4$	$0/2$	—

The roster has eleven models. A panel is unanimous when all valid votes agree, with at least four valid votes. Some panels have seven to ten valid votes rather than eleven; no item has fewer than four. "Disagreement" means disagreement with the computed or executed reference. When intent is withheld, another reading can be defensible.

Frozen result. The registered paired comparison is 0/146 on the stated side against 28/106 on the withheld side, a $+26.4$-point difference. Zero of the $20{,}000$ sampled sign assignments met or exceeded it; the plus-one Monte Carlo value is $1/20{,}001$, not zero. Constraint checking and SQL pass their registered family tests. Arithmetic is null. The observed stated-side zero is not a guarantee: zero of 124 all-class stated unanimous panels has an approximate one-sided 95% upper bound of 2.4%.

Author task-class ratings. Four paper authors rated eleven neutral-ID prompt classes. Agreement with the construction labels was 9/11 for the principal investigator (PI), 11/11, 9/11, and 8/11 for the three collaborators. These are author ratings, not independent external validation. The PI rated SQL-withheld as determinate and the second arithmetic placebo class as evaluative. Applying the registered exclusion leaves 142 blocks and a $+30.1$-point conditional contrast, but 0 SQL pairs. All retained disagreements are then in constraint checking, so the sensitivity does not confirm the effect in both named strong families.

Table 8: Registered comparison and retained sensitivities. The PI exclusion was registered; draw-zero selection was added after the result.

Analysis	Blocks	Stated side	Withheld side
Frozen last-valid primary	214	0/146	28/106
PI-disputed classes excluded	142	0/72	22/73
Draw zero only	214	0/148	24/98

Table 9: Four rules on the same registered paired study. The registered conditional contrast is the primary result; the rest are diagnostics on the same replies. Grouped tests treat a bug class or a whole family as the unit, and the author-class exclusion removes every SQL pair, so its remaining events are constraint-only.

Rule	Stated	Withheld	Difference	Inference as reported
Registered conditional	$0/146$	$28/106$	$+26.4$ pp	plus-one permutation $p = 1/20{,}001$ on 214 blocks
Registered, author-class exclusion	$0/72$	$22/73$	$+30.1$ pp	142 blocks; 0 SQL pairs remain
Draw-zero conditional	$0/148$	$24/98$	$+24.5$ pp	sensitivity to the extra draws, not the planned one-draw design
Complete panel, all eleven	$0/256$	$7/256$	3.27 pp blocks, 2.73 pp items, $[0.93, 5.61]$	$p = 0.0078125$ block-wise; $p = 0.25$ grouped by family
Complete panel, author-class exclusion	$0/184$	$1/184$	0.704 pp blocks	$p = 0.5$; one constraint event remains
Fixed retry completion	$0/256$	$12/256$	5.607 pp blocks	$p = 0.25$ grouped by family

Collection discrepancy and selectors. The collection requested one draw, but an inherited helper gave five draws to a hash-selected third of the items. The files contain 18,436 draw records for 7,920 model–item cells. There are no duplicate model–item–draw keys, and recorded gold labels match the generated items. The frozen code keeps the last-valid record per model–item cell; it does not count repeated draws as independent judges. Draw zero, first-valid, and first-completed-record analyses are post-hoc sensitivities. Tables 19 and 20 retain the invalid and complete-vote counts for both selectors.

Conditional and fixed-denominator results answer different questions. Beside the frozen conditional result 0/146 versus 28/106, the post-hoc draw-zero, all-eleven event count is $0/256$ versus $7/256$ on fixed matched items per arm. Those 256 items are arranged in 214 matched blocks, not 256 independent pairs. Table 9 places the registered result and these diagnostics in one display, with the block and item weightings kept apart. The item-weighted difference is 2.73 points; equal weighting of the 214 blocks gives 3.27 points. Thus $7/256$ is not another expression of the 3.27-point block-weighted contrast. An incomplete or non-unanimous panel has event indicator zero; it is not certified correct. Six of the seven withheld events are SQL false acceptances from one case/whitespace class, and one is a constraint false rejection. The blockwise $p = 0.0078125$ becomes $p = 0.25$ when SQL is grouped by bug class or when whole families are grouped. Under the PI exclusions all SQL pairs disappear, leaving one event and $p = 0.5$. These concentration checks limit the fixed-denominator result rather than replacing the registered conditional primary.

A.1 Fixed retry completion

The locked post-hoc rule keeps every valid original draw-zero vote and fills only an invalid first vote with its unique existing 2048-token retry. Missing or invalid retries remain missing. It collects no new response and does not select a favorable retry. It reconstructs an adaptive acquisition policy, not a randomized token-budget comparison.

Table 10: Fixed-denominator full-roster reference-disagreeing unanimity in a locked post-hoc sensitivity. Counts are events/all fixed matched items, not events/complete panels. All contrasts are withheld minus stated.

Policy / cohort	Stated	Withheld	Equal-block Δ (pp)	Block bootstrap 95% interval (pp)	Item-weighted Δ (pp)	Block p
Base / all	0/256	7/256	3.271	$[0.935, 5.607]$	2.734	0.0078125
Base / PI	0/184	1/184	0.704	$[0.000, 2.113]$	0.543	0.5
Hybrid / all	0/256	12/256	5.607	$[2.804, 8.879]$	4.688	2.44×10^{-4}
Hybrid / PI	0/184	6/184	4.225	$[1.408, 7.746]$	3.261	0.015625

Base uses original draw zero. Hybrid retains every valid base vote and fills only invalid cells with their unique existing valid 2048-token retry. The original primary is unchanged. The all cohort has 214 matched blocks and 256 items per arm; the frozen PI-excluded cohort has 142 blocks and 184 items per arm. "PI" denotes the previously fixed aggregate exclusion, not a newly selected population. Each block has equal weight; multiple variants within a block are averaged. The interval is for the equal-block contrast, *not* the item-weighted rate or contrast. An incomplete panel's zero full-event indicator does not certify correctness.
Pooling *only SQL* blocks by SQL bug class (leaving non-SQL registered blocks separate) gives full-cohort p =0.25 for base and p =0.0078125 for hybrid. Whole-family p remains 0.25 for all and 0.5 for PI; whole-family intervals are undefined. These results do not establish general robustness or a causal token-budget effect.

Table 11: Constraint-family coverage and unresolved cells. Panel fractions use all items in the arm. Repairs and failures count judge–item cells, not panels.

Arm	Base complete	Hybrid complete	Cells repaired	Cells unresolved	Content filter	Length
Stated	36/100	97/100	82	3	3	0
Withheld	4/100	38/100	141	65	58	7

The last two columns are the recorded retry finish reasons among unresolved cells. Cells with no existing retry: 0 stated and 0 withheld. Missingness therefore is not solely a length-limit issue; a missing vote is not an agreeing or correct vote.

On the fixed full cohort, the hybrid result is 0/256 versus 12/256. Under the PI exclusions it is 0/184 versus 6/184. Whole-family tests remain $p = 0.25$ overall and $p = 0.5$ after PI exclusion. Completion improves, but coverage remains strongly arm-associated; unresolved cells include both recorded content-filter and length outcomes. The rule does not make missingness ignorable or identify a causal budget effect.

Completing panels can also break agreement. Among withheld constraint panels with at least four valid votes, reference-disagreeing unanimity falls from $18/37$ to $14/32$ (reference-disagreeing/unanimous panels), even as full-roster events rise from 1 to 6 on the fixed 100-item arm. Re-asked votes break unanimity in 4 previously reference-disagreeing panels. These are distinct endpoints under the same policies, not a changed primary.

The older conditional retry calculation is different. It starts from the last record of any validity and then substitutes a valid retry only when that record is invalid. It can overwrite, change, or drop a vote available under the frozen last-valid selector. Its conditional contrast is therefore not a clean missing-vote repair under the same selection rule.

Table 12: Historical acquisition versus retrospective selected-record accounting, across all four families. Recorded `tok` is not a total input-token or monetary cost. New judging calls: 0.

Accounting category	Records	Recorded `tok`
Historical: all original draws	18436	3902205
Historical: all retries	404	336255
Hybrid: retained valid base votes	7600	1485471
Hybrid: selected valid retry votes	250	248742

The lower rows are selected subsets, *not* additional acquisition and not an accounting of all failed attempts. Including invalid original draw-zero records and available retry attempts gives 8240 records and 1992182 recorded `tok`. These are retrospective records, not a prospective matched-cost comparison.

The separate legacy sensitivity uses the *last record of any validity*, then a valid retry only when that last vote is literal `?`. It makes 217 constraint and 23 SQL substitutions; constraint conditional counts are 0/15 stated versus 20/38 withheld. Pooled, it gives 26/103 minus 0/144 $=$ 25.243 pp across 214 pairs: 0 exceedances/20000, plus-one $p = 1/20001$, never zero. Of its constraint substitutions, 40 overwrite a vote already retained by the registered last-valid selector (4 change the vote), and 2 registered votes are dropped. Its 217 substitutions are not the hybrid's 223 draw-zero constraint repairs.

The twins share the complete stored record and reference label CORRECT. All 30 stored strict-constraint checks pass. One field of that shared record is {"bin_28": "HXKMA6T"}; this is a record excerpt, not the entire record.

Base task	One actual task-statement line
Stated	C01: bin_28 must start with the exact prefix "HX".
Withheld	C01: bin_28 must use the standard prefix for this record type.

Excerpt of a whole-specification change: the header and all 30 constraint lines change wording. The displayed line is not a single missing-rule intervention. These quotations come from the stored *base-task statements*, which ask for VALID/INVALID; they are not quotations from a common judging wrapper. The wrapper is not reproduced here. The verdicts below are the stored CORRECT/BUGGY judge labels under the original last-valid selector.

Arm	CORRECT	BUGGY	Unparsed	Valid/roster
Stated	8	3	0	11/11
Withheld	0	10	1	10/11

"Unparsed" here means no valid vote was retained from any original draw; it is not another agreeing vote. The withheld panel is unanimous only among its 10 valid votes, not all 11 roster members. The stated panel is complete but not unanimous.

Figure 2: An outcome-selected illustration, not a representative item or a full-eleven-vote unanimity example. Selection used the original frozen last-valid reference-disagreeing withheld panel, before the hybrid analysis; the shared reference is checked against the strict specification.

A.2 Inspectable constraint example

Section 4 summarizes this stored pair; the pair itself is Figure 2. It was selected by the lowest pair identifier meeting the frozen last-valid, unanimous-withheld reference-disagreement condition, before any hybrid outcome was examined, so it is an illustration and not a sampled item.

The example shows an excerpt from a whole-specification rewrite, not a single omitted rule. Appendix F gives the full input template and the frozen collector's direct-message path; no separate common judging wrapper is added. The withheld wording does not uniquely determine the executed strict predicates, and the valid-vote counts must not be described as an eleven-way decision.

A.3 Retained main-text details

An earlier cohort, where the requirement is withheld rather than reversed. The registered paired study asks a weaker version of the same question: its withheld arm leaves the convention unsaid instead of stating its opposite, its panels are eleven judges with missing votes, and its decision rule is conditional on unanimity. Its stored twins change the whole specification – the header and all 30 constraint definitions are rewritten, not one clause removed – and the withheld panel it produced is unanimous only among its valid votes, a cell with no parsed reply being a missing vote rather than another agreeing one. Unlike the policy-slot test, this constraint comparison has no length-matched placebo; the candidate and designated reference stay fixed within each pair. Figure 2 in Appendix A.2 keeps that pair with its vote counts. We report the registered result as registered, with the diagnostics that qualify it.

Registered result. On 214 eligible paired blocks, the frozen analysis gives $0/146$ reference-disagreeing unanimous panels on the stated side, including placebo, versus $28/106$ on the withheld side. The difference is $+26.4$ percentage points (plus-one permutation $p = 1/20{,}001$). Constraint checking contributes 22 disagreements and SQL contributes six; paired arithmetic contributes none. Four authors rated the eleven task classes. The principal investigator agreed with construction on 9/11 classes and classified SQL-withheld as determinate. The registered exclusion of disputed classes leaves $+30.1$ points over 142 blocks, but removes all SQL pairs. All retained errors then come from constraint checking. This is a constraint-driven sensitivity, not independent human confirmation across both positive families. Appendix A gives the family table, frozen criteria, and valid-vote distributions.

Collection discrepancy and complete-panel check. The intended collection requested `--k 1`, but an inherited helper sampled a hash-selected third of items five times, so the archive has 18,436 draw records for 7,920 model–item cells. The frozen selector takes the last valid record in a cell; a later draw-zero sensitivity retains a $+24.5$-point conditional contrast without turning the collected data into the planned one-draw experiment. Requiring all eleven draw-zero votes gives $0/256$ stated versus $7/256$ withheld reference-disagreeing unanimous events. Six are SQL false acceptances, all from one case/whitespace class; one is a constraint false rejection. The blockwise $p = 0.0078125$ becomes $p = 0.25$ under bug-class or family grouping, and excluding the disputed classes leaves one event ($p = 0.5$). Completion is arm-associated too: only $4/100$ withheld constraint panels are complete, against $36/100$ stated panels. The item-level contrast is therefore not evidence of a broadly replicated task-family effect. A separately fixed retry rule, which fills only invalid first votes from already collected 2048-token replies, gives $0/256$ versus $12/256$ complete-panel events ($p = 0.25$ grouped by family), and $0/184$ versus $6/184$ after the author-class exclusion ($p = 0.5$); coverage stays unequal and no token-budget effect is isolated. Table 9 in Appendix A sets these endpoints beside the registered result, with the block and item weightings kept apart (Appendix A.1).

A second task domain, with a registered null. We also tested an answer convention on public FOLIO formal premises (Han et al., 2024). The registered all-eleven unanimity endpoint is $0/263$ in both arms: the registered test found nothing. A later post-hoc analysis of the same replies finds $0/263$ versus $105/263$ disagreements at six votes, and 0 versus 28 at seven votes, while every automatic decision still requires all eleven replies recognized. Coverage moves with the threshold: the six-vote rule decides 258 specified and 255 omitted-convention panels, whereas unanimity decides no omitted-convention panels at all. When the convention is absent, `FALSE` can reasonably mean "not entailed" rather than "refuted," so these are convention-relative policy results inside a family of 54 comparisons, not a positive replacement for the registered null; re-coding reasons under a changed binary reference is a different convention rather than an improvement under the unchanged three-way truth, and the one-vote bound of 28 to 105 is arithmetic rather than an experiment. Appendix E.2 retains every cutoff, coverage, control and interpretation-coding limit, with the registered endpoint and the threshold policies in one display (Table 28).

B Fresh matched-policy studies

These studies were fixed prospectively after an exploratory analysis of earlier responses. They use fresh cases and policy wordings, not a relabeling of the earlier registered endpoints. All original replies and nonvalid outcomes are retained. The four judges are GPT-6 Astra, GPT-5.6, Opus 5 and Opus 5.5, spanning two providers. They are a fixed roster, not independent task mechanisms.

B.1 Corpora and controlled comparisons

Both studies use eight authored generators: interval endpoints, name matching, set versus multiset intersection, rounding ties, timestamp dates, graph direction, sorting ties and empty-group averages. Two executable references evaluate each candidate under the two specified conventions. Ambiguous cases have opposite candidate verdicts; determinate controls have the same verdict. Fresh exact case data and prompt hashes do not overlap the earlier corpus or the other new cohort. The generators remain the same families; this is not replication on newly sampled mechanism families.

The first study contains 160 cases: 128 ambiguous and 32 determinate, balanced within mechanism by designated convention and candidate correctness. Each case receives three independent fresh-context calls per judge. Specified (S) and paraphrase (P) state the same policy; opposite (A) changes the policy and the applicable reference. Task, input, candidate and header are fixed within case. The policy slots each have 12 whitespace-separated words and matched ASCII character length. In these clauses, paraphrases have 11–12 word edits and opposite rules 1–8; these are not model-token distances or elimination of all lexical cues. The 1,920 assigned calls include all failures.

The second study contains 96 disjoint cases: 64 ambiguous and 32 determinate. Each prompt supplies the same exhaustive two-policy menu and states that exactly one policy governs a task requiring one result. The joint-assessment baseline receives no selected policy and is explicitly instructed to assess both completions. It is therefore already policy-aware. The separate-policy procedure selects policy 0 in one call and policy 1 in another. Both procedures have two planned fresh-context calls

Mechanism	S/A changes	S/P changes	Complete triplets	Bounds (pp)
Name matching	55	4	55	[68.75, 89.06]
Empty-group averages	64	0	64	[100.00, 100.00]
Graph direction	61	0	61	[90.62, 100.00]
Intersection multiplicity	61	0	61	[92.19, 100.00]
Interval endpoints	56	0	56	[78.12, 100.00]
Rounding ties	56	1	59	[81.25, 90.62]
Sorting ties	52	0	52	[65.62, 100.00]
Timestamp dates	61	0	61	[92.19, 100.00]

Table 13: Fresh explicit-policy study. Each mechanism has 64 assigned ambiguous judge-case profiles. Change counts are observed events over those assignments; complete-triplet counts state availability, not another source of votes. The final column bounds the assigned S/A-minus-S/P contrast under all allowed completions of nonvalid labels.

per judge-case and the same output contract and aggregation: retain an agreeing pair of valid labels, return `UNDETERMINED` when valid labels differ, and mark the pair nonvalid if either call is nonvalid. Nonvalid is not abstention. This yields 1,536 planned calls. The comparison tests decomposition with a known menu, not discovery of arbitrary missing policies. Executable reference outputs and oracle verdict labels are not model inputs.

B.2 Readout, acquisition and analysis

The decision extractor was fixed before collection. It accepts exactly one complete JSON object with the keys `verdict` and `reason`, a permitted verdict (`CORRECT`, `INCORRECT`, `UNDETERMINED`) and a string reason. Extra prose is separately recorded as format noncompliance, not silently treated as a wrong decision. Multiple qualifying objects or qualifying objects nested inside wrappers are invalid. Reasons longer than 25 words remain verdict-valid and are flagged separately. An exact-format-only readout was also fixed before the first study's calls as a same-response sensitivity. It does not change any earlier study's scores.

All calls use a 1,024 completion/reasoning-token cap. OpenAI routes use low reasoning effort; Opus 5 disables thinking and Opus 5.5 uses adaptive low effort. Temperature, top-p and sampling seed are not explicitly set; provider defaults are not claimed to be equivalent. Returned model identities are checked. The two baseline calls are separate contexts, not a claim of stochastic independence. Only pre-content connection failures and specified transient HTTP failures permit one identical retry. Partial HTTP-200 streams, truncated responses, parser failures and unfavorable verdicts are not rerolled. There are 3,459 attempts for 3,456 assigned calls: 3,325 valid replies, 67 parser failures and 64 acquisition failures.

The first primary contrast is the observed assigned-profile S/A verdict-change rate minus the S/P change rate over 512 ambiguous judge-case profiles. Complete-triplet results use the same three-response cohort for both changes. Exact completion bounds allow every nonvalid label any permitted verdict, keeping observed valid labels fixed; they neither impute votes nor constitute confidence intervals. The second primary is separate-policy minus baseline operational accuracy, weighting ambiguous and determinate cases equally within each mechanism, then judges and the eight mechanisms equally. A nonvalid pair fails its target. We report mechanism-bootstrap percentile intervals from 10,000 draws and leave-one-mechanism-out summaries as finite-benchmark stability, not population coverage or an equivalence test. No effect-based extension or positive-result stopping threshold is used.

B.3 All mechanisms and complete-panel outcomes

Across the same 469 complete triplets, S/A changes 462 verdicts and S/P changes 4. Across all 512 assignments, the observed-event contrast is 90.04 percentage points, with completion bounds $[83.59, 97.46]$. Joint reference correctness is $466/512$ for S/A and $462/512$ for S/P/A. The strict-format sensitivity retains a 73.44-point lower bound. In determinate controls, S/A retains the verdict on $119/120$ valid pairs and S/P on $118/118$ valid pairs. All-four S/P/A reference correctness is

Mechanism	Joint baseline (%)	Separate policies (%)	Difference (pp)	Bounds (pp)
Name matching	95.31	95.31	0.00	[-4.69, 3.12]
Empty-group averages	96.88	100.00	3.12	[0.00, 3.12]
Graph direction	100.00	100.00	0.00	[0.00, 0.00]
Intersection multiplicity	100.00	98.44	-1.56	[-1.56, 0.00]
Interval endpoints	98.44	98.44	0.00	[-1.56, 1.56]
Rounding ties	90.62	92.19	1.56	[-6.25, 4.69]
Sorting ties	75.00	73.44	-1.56	[-26.56, 25.00]
Timestamp dates	98.44	96.88	-1.56	[-3.12, 1.56]

Table 14: Same-menu two-call comparison. Accuracy gives the two case types equal weight, retaining nonvalid pairs as failures. Bounds conservatively complete nonvalid pairs; they are not population confidence intervals.

Procedure	Planned calls	Attempts	Measured input tokens	Measured output tokens	Unknown usage
Joint baseline	768	771	227,025	71,216	30
Separate policies	768	768	226,380	66,400	28

Table 15: Observed acquisition workload for the same-menu study. Measured token totals exclude attempts with unknown usage; unknown usage is retained in conservative accounting. Equal planned calls and caps do not imply equal tokens, realized attempts or deployment cost. No cost-saving claim follows.

$78/128$ assigned ambiguous cases, with 85 complete three-arm rosters. This is distinct from the judge-case change rates. Missing judges are never agreeing votes.

Both procedures obtain $241/256$ correct ambiguous deferrals and $121/128$ correct determinate decisions, or 94.34% balanced accuracy. The observed difference is 0.00 percentage points; its mechanism-bootstrap interval is $[-0.98, 1.17]$ and completion bounds are $[-5.47, 4.88]$. The result establishes neither superiority nor equivalence.

Separate-policy judgments both match their respective references on $237/256$ ambiguous profiles. 4 other correct aggregate deferrals involve at least one conditional abstention. Correct deferral and two reference-aligned conditional judgments are therefore different observations. The baseline's prose was not coded for conditional information, and downstream evaluator benefit or internal reasoning was not measured. Baseline repeat labels differ on only $2/363$ valid pairs (two of 242 ambiguous, zero of 121 determinate pairs): the nominal two-draw baseline behaved almost deterministically. Disagreement between opposing selected-policy calls is not a comparable repeat-noise measure.

The evidence concerns the fixed four-judge roster, authored mechanisms and specified policy menu. It does not supply new independent human coding, naturalistic prevalence, or verification beyond the chosen executable references. The accompanying anonymous supplemental ZIP contains these cases, exact prompts, saved response texts, attempt statuses and offline analysis with hash-bound provenance (Appendix H); no public release link is asserted.

C Disclosure measures and human/probe controls

The disclosure results use several instruments. A coded rationale, a menu token, and an answer to a direct follow-up are not interchangeable. None establishes what a judge represented before it was asked.

C.1 Two-author rationale coding

Two paper authors, the PI and one collaborator, completed the same 200-rationale sample. The original model-coded pool had $2{,}095$ rationales; six were reserved as instruction examples, leaving a $2{,}089$-rationale frame. The sample includes all 52 remaining unanimous-model-flag rows, all 25

Table 16: Disclosure estimates on the human-coding frame. The model majority is a comparison, not ground truth. Human union is the registered primary.

Rule	Ambiguous	Explicit	Placebo	Agreement with model
Model majority	6.5%	0.6%	0.8%	—
Human union	16.6%	5.4%	12.7%	91.5%
Human intersection	5.7%	0.9%	1.0%	99.1%
PI alone	13.6%	5.4%	12.5%	91.7%
Collaborator alone	8.8%	0.9%	1.1%	98.9%

split-model rows, and 123 of 2,012 unanimous-model-nonflag rows. The reported frame rates use the frozen inverse-probability weights. Raw sample percentages are not frame estimates.

Both coders are authors. Their sheets withheld arm, reference, judge identity, and prior model code, but complete independence from the hypothesis or wider study outcomes is not established. Raw agreement is 87.0%, and nominal $\alpha = 0.715$ passes the registered sample-level reliability gate. With two coders, the registered primary is the union: either author's flag counts. Intersection and individual-coder estimates remain sensitivities.

The union ambiguous estimate is 16.6% ([9.3,25.6]%). Its difference from explicit is $+11.2$ points ([1.8,21.1]), while the $+3.9$-point placebo difference has interval [-11.7,17.6] and is unresolved. Positive-class agreement is limited: union-versus-model Jaccard is 0.251, and frame-weighted human–human Jaccard is 0.239. The frozen outcome is SUPPORTS-REVISED, not full support or numerical reproduction of the model estimate.

This sample contains rationales, not complete eleven-judge panels. It cannot recompute or validate complete-panel silence, and it does not validate the later menu or follow-up probes. Those endpoints remain identified as model-coded or parsed-token measurements.

C.2 Manipulation and model-coding controls

Two models outside the twelve-judge panel and outside its vendor families rated the paired wording without the candidate, reference, or arm label. On the fresh cohort, the primary rater called $0/65$ explicit and $60/61$ ambiguous questions under-specified; the second gave $2/65$ and $39/61$. Both called a length-matched placebo determined ($0/32$). These are model ratings, not independent human validation. They support that the wording manipulation changes whether the intended axis is stated, while leaving the reference-choice problem described in Appendix F.3.

The original spontaneous-disclosure protocol used three model coders, blind to arm, reference, verdict, and item, on 2,095 rationales. A flag names the missing requirement, states an adopted assumption, or asks for clarification. The original round had $\alpha = 0.86$; positive-class pairwise agreement was 0.74, 0.74, and 0.81. The baseline rate used with the menu arms comes from a separate pooled re-coding round: 1,570 baseline and 1,565 permitted-arm rationales, with $\alpha = 0.847$. It is not a superset of the human-coding frame.

C.3 Changing the response menu

The menu experiment used the same 144 SQL items and eleven-judge panel. Its design and numerical endpoints were frozen before collection; interpretation of overlapping outcome branches was partly post-hoc. Baseline and permitted signals are model-coded rationale labels. Offered and direct signals are parsed verdict tokens, and direct removes the candidate. These instrument differences limit cross-arm comparisons.

Merely offering an UNDERSPECIFIED verdict leaves panel silence nearly unchanged. Inviting the judge to state that the task is open reduces it, but 51.4%, with interval $[37.5, 65.3]\%$, misses the registered below-0.50 point-estimate recovery branch by 1.4 points. Among 311 cells where the direct arm says OPEN, the same judge names the issue while judging in 44 (14.1%). This does not show suppression: the direct prompt can construct or elicit a distinction that was not represented at verdict time.

Table 17: Disclosure under four prompts. Panel silence is the fraction of ambiguous items on which no judge flags.

Endpoint	Baseline	Offered	Permitted	Direct
Ambiguous judgments flagging	6.23%	11.24%	15.35%	40.03%
Explicit-control flagging	0.77%	0.13%	0.89%	1.52%
Ambiguous-panel silence	72.2%	70.8%	51.4%	22.2%

The accuracy and silence observations come from different arms. Offered shows binary accuracy on its remaining votes moving from 0.686 to 0.720 on ambiguous items; permitted supplies the larger silence reduction. The denominator changes in offered, half the panels remain silent, and the pair is a direction for study rather than a validated remedy. On the confidently wrong stratum, all four flag-enrichment ratios are at or below 1.05 and none is significant.

C.4 Same-conversation post-verdict probe

One conversation was run for each of eleven judges and 144 items. Turn 1 was byte-identical to the ordinary baseline; turn 2, after the verdict and with the candidate still in context, asked whether the wording settled the required behavior. There were $1{,}584$ conversations and $1{,}561$ turn-2 responses; 23 unparsed turn-2 responses remain non-disclosures.

On all 792 under-specified cells, 333 probes return OPEN (42.1%). The paired comparison uses the 780 cells with a parsed verdict and probe: 275 are silent at turn 1 and open at turn 2, against one reverse. All 36 base questions move in the same direction. On explicit controls the probe returns open on $65/792$ cells (8.2%). Panel silence falls from 72.2% to 12.5%. The explicit control rules out indiscriminate acquiescence, but the result remains an elicitation effect, not evidence of conscious concealment or prior awareness.

C.5 Two-turn regex replication

The same pre-registered two-turn design was run on 144 regular-expression items. Among $1{,}398$ usable answers, the ambiguous arm is open on $338/701$ cells and the explicit control on $46/697$, a control-corrected $+41.6$ points $[+35.2, +47.8]$. Every fully collected judge and all six ambiguity classes move in the same direction.

There are 186 unusable cells, mostly from one judge's turn-1 content-filter failure. Scoring every missing ambiguous cell as determined and every missing explicit cell as open leaves $+24.9$ points. This bound avoids assuming that missingness is ignorable. Registered secondary comparisons were not completed, so the claim is the disclosure contrast on a second substrate, not completion of the full registered plan.

C.6 One-turn regex control defect

The separate one-turn regex test has a different endpoint. Direct returns open on 64.3% of ambiguous cells and offered returns UNDERSPECIFIED on 14.4%, so the raw registered difference is $+49.9$ points and passes its registered branch. The direct probe also calls 32.3% of explicit and 27.0% of placebo cells open. Control correction reduces the explicit comparison to $+17.9$ points $[10.2, 25.3]$, below the registered 20-point threshold; placebo correction gives $+23.4$ points.

Review of the flagged control rationales found 142 references to a second ambiguity axis left open by the supposedly explicit wording. The generator checked that the candidates differed, not that every other interpretation was settled. The raw registered result and the control-adjusted limitation are both retained. The absolute disclosure rate is prompt-dependent and must not be combined with the two-turn result as though they used one instrument.

Table 18: Measurement crosswalk. Similar rates come from different records and outcome definitions.

Measurement	Unit	Outcome
Spontaneous SQL disclosure	Rationale	Three-model code under a frozen rubric
Two-author check	Sampled rationale	Frozen human-union primary and sensitivities
Menu experiment	Judge–item cell	Model code or parsed menu token, depending on arm
SQL post-verdict	Conversation	Model-coded turn 1 and parsed OPEN at turn 2
Regex two-turn	Conversation	Parsed OPEN, with adverse missingness bound
Regex one-turn	Judge–item cell	Parsed direct/offered tokens with explicit/placebo controls

D INTERVENTION OUTCOMES AND COSTS

The interventions below use different cohorts and outcomes. A lower flag rate, more complete voting, or a changed threshold is not by itself evidence of better decisions. Each result therefore reports its eligibility rule, both error directions, and the available cost or coverage comparison.

D.1 OPERATIONAL CHECKS ON THE FRESH RECORDS

The base study files contain $18{,}436$ draw records for $7{,}920$ model–item cells. Separate higher-budget retries contain 404 records, of which 332 parse as valid votes. The draw-zero and frozen last-valid selectors in Tables 19 and 20 do not merge those retries. The retry-completion rule is reported separately in Appendix A.1.

The fixed-denominator comparison uses 256 stated and 256 withheld items arranged in 214 matched blocks. These are fixed matched items per arm, not 256 independent paired blocks. As reported beside the registered conditional primary in Appendix A, draw zero with all eleven votes gives $0/256$ versus $7/256$. The item-weighted difference is 2.73 points; the equal-block contrast is 3.27 points. The event fraction $7/256$ and the 3.27-point block-weighted contrast are not interchangeable. An incomplete or non-unanimous panel has no automatic-decision event; that zero event indicator is not a correct decision. The six SQL false acceptances all come from one case/whitespace class, and PI exclusion removes every SQL pair.

Completeness changes both errors and useful decisions. On withheld draw-zero items, requiring all eleven votes rather than at least four reduces unanimous automatic decisions from 101 to 64 and reference disagreements from 24 to seven. It defers 17 wrong and 20 correct decisions. If only unanimous-CORRECT panels ship, the same six SQL false acceptances remain among the same 20 shipped items, while stated-side correct shipments fall from 62 to 49. More complete voting therefore reduces rejection-side events without removing the observed bad shipments.

Retrospective routing controls. The routing comparison uses 144 post-verdict SQL items and 36 base questions split into two folds. The loss is ten times a shipped reference-buggy item plus one times a deferred reference-correct item at the primary $k = 10$; it is not measured human-review error or total monetary cost. Both flag policies lose to the zero-inference defer-all control before charging for the extra inference: 53.5 and 60.4 versus 50.0 loss per 100.

Against calibrated confidence on the fixed split, the gate-only rule changes ten routing decisions favorably and none adversely, but the changes span seven base questions and three bug classes. A nested cluster bootstrap that refits thresholds gives interval [-50.69,7.66], crossing zero. The gate also fails to beat the better dissent comparator. These are retrospective held-out comparisons on a previously explored cohort, not prospective or cost-matched deployment results.

Table 19: Draw zero completeness by family/arm. N: items; raw/bad: all draw records/unparsed records before selection; missing: unparsed selected model–item cells; $v : n$: n items have v valid votes. U/E: unanimous panels/reference disagreements with at least four valid votes or all eleven. Raw records are the same base collection in both tables, not additional data; separate higher-budget retries are excluded from both selectors.

Family/arm	N	Raw/bad	Missing	$v : n$	$U_{\geq 4}/E$	U_{11}/E
Constraint/strict	100	2596/195	85	8:2, 9:17, 10:45, 11:36	18/0	2/0
Constraint/vague	100	2156/395	206	7:3, 8:28, 9:45, 10:20, 11:4	37/18	1/1
SQL/ambiguous	72	1584/19	9	9:1, 10:7, 11:64	32/6	32/6
SQL/explicit	72	1760/27	16	10:16, 11:56	44/0	43/0
SQL/placebo	48	1012/5	3	10:3, 11:45	30/0	29/0
Arithmetic 1/ambiguous	84	2288/1	1	10:1, 11:83	29/0	28/0
Arithmetic 1/explicit	84	2332/0	0	11:84	56/0	56/0
Arithmetic 2/placebo	32	880/0	0	11:32	4/0	4/0
Arithmetic 2/stated-a	32	924/0	0	11:32	7/0	7/0
Arithmetic 2/stated-b	32	968/0	0	11:32	2/0	2/0
Arithmetic 2/unstated	64	1936/0	0	11:64	3/0	3/0

Table 20: Frozen last-valid completeness by family/arm. N: items; raw/bad: all draw records/unparsed records before selection; missing: unparsed selected model–item cells; $v : n$: n items have v valid votes. U/E: unanimous panels/reference disagreements with at least four valid votes or all eleven. Raw records are the same base collection in both tables, not additional data; separate higher-budget retries are excluded from both selectors.

Family/arm	N	Raw/bad	Missing	$v : n$	$U_{\geq 4}/E$	U_{11}/E
Constraint/strict	100	2596/195	68	9:13, 10:42, 11:45	16/0	3/0
Constraint/vague	100	2156/395	173	7:3, 8:19, 9:40, 10:24, 11:14	41/22	6/5
SQL/ambiguous	72	1584/19	7	9:1, 10:5, 11:66	33/6	33/6
SQL/explicit	72	1760/27	10	10:10, 11:62	44/0	43/0
SQL/placebo	48	1012/5	2	10:2, 11:46	30/0	29/0
Arithmetic 1/ambiguous	84	2288/1	0	11:84	32/0	32/0
Arithmetic 1/explicit	84	2332/0	0	11:84	56/0	56/0
Arithmetic 2/placebo	32	880/0	0	11:32	4/0	4/0
Arithmetic 2/stated-a	32	924/0	0	11:32	7/0	7/0
Arithmetic 2/stated-b	32	968/0	0	11:32	1/0	1/0
Arithmetic 2/unstated	64	1936/0	0	11:64	2/0	2/0

D.2 Stating the missing invariant

The antecedent twelve-judge SQL study uses 28 reference-correct non-canonical variants that pass 5,600 differential executions without a discrepancy. Most judges nevertheless reject them more than half the time; majority-vote error is 0.875. Mean pairwise error correlation is -0.003 with interval $[-0.05, +0.05]$, while independent-error calculations at the observed marginals give 0.864 and 0.889. This is compatible with voting amplifying a shared marginal bias; pairwise decorrelation does not establish joint independence. Execution equivalence is relative to the designated reference, whose empty-set behavior the request does not fully state.

A pre-registered five-arm experiment used 57 items and twelve judges. Baseline is the original prompt. Ruled-out states that every parent has at least one child; asserted states the opposite; placebo adds matched irrelevant text; the fifth arm directly discloses the reference behavior. False-alarm rates on reference-correct items are 0.622 at baseline, 0.107 under ruled-out, 0.744 under asserted, and 0.717 under placebo. Panel error moves $0.804 \to 0.000$ under ruled-out but to 1.000 under placebo. The two wrong-way controls prevent reading extra text alone as the repair.

The result localizes an under-specified empty-set choice; it does not establish a second mechanism or a general clarification policy. A separate deployment generator began with 160 items and retained 155 after a reference-based label-stability screen. On the 80 retained reference-correct items, tie-scored panel errors fell from 23.5 to zero (a tied vote counts as half an error). Neither arm missed any of the 75 retained buggy items.

Five additional wrong-join candidates became equivalent to their references under the stated invariant and were excluded before judging. They are not five observed new misses, so dividing the removed errors by five does not estimate a harm threshold. The screen itself uses references and wit-

Table 21: Post-hoc fixed-denominator joint events. Counts use items; differences and intervals weight matched blocks equally. Missing/non-unanimous panels make no automatic-error event, not a correct decision. Placebo is excluded.

Rule	Blocks	Stated	Withheld	Gap (pp)	95% CI	p
Draw zero, all 11	214	0/256	7/256	3.27	$[0.93, 5.61]$	0.007812
Draw zero, ≥ 4 valid	214	0/256	24/256	11.21	$[7.01, 15.42]$	5.96×10^{-8}
Draw zero, all 11, PI exclusion	142	0/184	1/184	0.70	$[0.00, 2.11]$	0.5
Last valid, all 11	214	0/256	11/256	5.14	$[2.34, 8.41]$	4.883×10^{-4}
Last valid, ≥ 4 valid	214	0/256	28/256	13.08	$[8.88, 17.76]$	3.725×10^{-9}
Last valid, all 11, PI exclusion	142	0/184	5/184	3.52	$[0.70, 7.04]$	0.03125

Table 22: Retrospective two-fold evaluation on 144 SQL items at $k = 10$. Bad ship and correct defer are reference-defined. Review includes all deferred items, not measured human outcomes. Turns are nominal inference turns per item; calibration, token, retry and review costs are not fully matched.

Policy	Ship	Bad ship	Correct defer	Review	Loss/100	Turns
Ship all	144	72	0	0	500.0	0
Defer all	0	0	72	144	50.0	0
Majority	69	9	12	75	70.8	11
Unanimous acceptance	30	6	48	114	75.0	11
Calibrated dissent	39	3	36	105	45.8	11
Calibrated margin	59	9	22	85	77.8	11
Calibrated confidence	35	7	44	109	79.2	11
Flag gate only	39	4	37	105	53.5	22
Gate plus automatic rescue	40	5	37	104	60.4	22
Selected single judge	72	12	12	72	91.7	1
Single-judge confidence	37	6	41	107	70.1	1
Test-oracle single (non-deployable)	67	7	12	77	56.9	1

nesses. A gold-free pre-flight intended to predict when disclosure was needed failed its registered gates (specificity 0.000, AUROC 0.000, and a wrong-signed rank correlation). The measured gain is therefore for a screened cohort with a verified precondition.

D.3 Decision value of the elicited flag

The pre-registered decision-value analysis ranks the same 144 SQL items at matched coverage; it makes no new model calls. The post-hoc error-channel split shows the flag on all $11/11$ false-alarm items but only $3/9$ missed-bug items. The corresponding mean panel flag rates are 0.727 and 0.354; the latter is indistinguishable from the 0.367 rate on under-specified items the panel got right. The split rests on only 20 panel errors.

The registered pooled criterion is not fully met. The flag beats vote margin, but not mean stated confidence ($p = 0.059$). At 50% coverage it accepts fewer missed bugs than margin or random escalation, and there is no cost crossover for $k \in [1, 50]$. Per judge it can lose to the judge's own confidence. Only 12 of 36 base questions carry an error, all in two of six bug classes, and reducing the under-specified share from 50% to 10% removes the pooled advantage.

At ten-to-one cost, the nine missed bugs and eleven false alarms contribute 101 units, only 11 of them from false alarms. Even a perfect false-alarm-only instrument could recover about 11% of this pool's cost. That is not a transferable ceiling. The flag is useful here as a scoped false-alarm warning, not as a general bug detector or a rule for distrusting individual judges.

D.4 Verification and its two error channels

On 138 of 224 canonical variants complete in both arms (828 of $1{,}344$ possible critic cells from 84 of 112 tasks), aggregate per-critic error falls from 15.8% to 7.1%. The direction differs by class: clean false accusations fall $117/366 \to 0/366$, while buggy misses rise $14/462 \to 59/462$ (3.0% to 12.8%). The clean-side zero is a same-reference ceiling because tests are retained only when they pass that clean program. It is not a general correctness guarantee.

The consensus guard is a separate paired comparison on 642 model-generated, visible-test-passing implementations in 112 tasks. OR-aggregated flags fall $142/642 \to 41/642$ (22.1% to 6.4%). Of

Table 23: Routing-loss sensitivity. Each calibrated policy is fitted for that column's k on opposite-fold labels. This is not a post-hoc choice of the best cost ratio.

Policy	$k = 1$	$k = 5$	$k = 10$	$k = 25$
Ship all	50.0	250.0	500.0	1250.0
Defer all	50.0	50.0	50.0	50.0
Majority	14.6	39.6	70.8	164.6
Unanimous acceptance	37.5	54.2	75.0	137.5
Calibrated dissent	27.1	35.4	45.8	77.1
Calibrated margin	14.6	46.5	77.8	171.5
Calibrated confidence	11.8	41.7	79.2	66.7
Flag gate only	12.5	29.2	53.5	54.9
Gate plus automatic rescue	10.4	32.6	60.4	72.2
Selected single judge	17.4	50.0	91.7	216.7
Single-judge confidence	17.4	48.6	70.1	60.4
Test-oracle single (non-deployable)	11.1	32.6	56.9	129.9

the 101 removed flags, 58 become CORRECT and 43 abstain. Coverage is $599/642 = 93.3\%$. On a separate bed of 112 certified buggy mutations, catches fall $104/112 \to 100/112$ (92.9% to 89.3%): three become abstentions and one is accepted.

This is a plain-verification-to-guard comparison. There is no matched bare-vote baseline on the 642 implementations, and the guard needs external references and local execution. Its lower flag rate must be reported with abstention, coverage, and lost catches.

E A PUBLIC-FORMAL BOUNDARY TEST AND RECORDED-VOTE POLICIES

E.1 REGISTERED FOLIO COMPARISON

Source and manipulation. The study uses the public FOLIO v0.0 train and validation release at repository revision `5d7bb84c7edab3fb358e057d2807f19cf5cf5e2d`. The selected 263 blocks have neutral conclusions: checked finite models of the formal premises give both truth values. This verifies neutrality under the formal interpretation, not English–formula fidelity or accuracy on later benchmark labels. Selection, exclusions, and duplicate removal preceded judging. The blocks contain 230 coarse shapes and 236 structural signatures, not 263 independent reasoning mechanisms.

The candidate answer is `FALSE`. Under the stated three-way convention, `TRUE` means entailed, `FALSE` means refuted, and `UNKNOWN` means neither, so the candidate's designated reference verdict is INCORRECT. Without the convention, the binary reading "not entailed" is plausible. The two primary arms add either the convention or a matched neutral paragraph to the same premises, conclusion, and candidate. They measure disagreement with the stipulated three-way reference, not error under every plausible convention.

An event requires all eleven scheduled first replies to be returned, recognized, identical, and reference-disagreeing. Missing, unrecognized, or abstaining replies prevent an event; they are not imputed as correct. The fixed denominator is 263 blocks in each arm. Invariant-INCORRECT and invariant-CORRECT controls were assigned in both arms.

Fixed panel and collection. The panel is Claude Haiku 4.5; Claude Opus 4.6, 4.7, 4.8, and 5; Claude Sonnet 4.6 and 5; GPT-5.4 Mini, GPT-5.5, and GPT-5.6; and the requested MAI Code Flash alias. Ten judges used the direct transport. The unavailable GPT-5.5 route was replaced prospectively by an isolated Copilot-hosted GPT-5.5 request with the same supplied task text; the interface difference is retained. The MAI alias reported a different backend name and was recorded rather than silently treated as the requested deployment.

All 21,703 scheduled jobs reached a terminal state. There are 21,699 returned replies, 62 unparsed first replies, and four primary jobs with no returned answer, including three terminal authentication failures. Returned answers were not rerolled for label or format. The strict parser and an independent recomputation have zero recognition disagreements over the returned scientific replies.

The registered result is 0/263 versus 0/263, a 0.0-point difference with one-sided exact $p = 1.0$. The technical completeness gate passes. The bootstrap interval is $[0, 0]$ because every observed event

Table 24: Registered full-panel endpoint. Every event requires all eleven recognized C/I verdicts. Completeness is a technical count, not an outcome.

Probe type	Specified convention	Matched placebo
Neutral reference-disagreeing unanimity	0/263	0/263
Invariant-INCORRECT control	$0/263$	$0/263$
Invariant-CORRECT control	$0/66$	$0/66$
Fully recognized neutral panels	258/263	255/263

Table 25: Post-hoc recorded-vote policies on neutral probes. S is specified; P is the matched placebo. Conditional disagreement is undefined when a policy makes no automatic decision.

	Reference disagreements		Automatic decisions		Conditional disagreement	
k	S	P	S	P	S	P
6	0/263	105/263	258/263	255/263	0/258	105/255
7	0/263	28/263	258/263	102/263	0/258	28/102
8	0/263	2/263	258/263	25/263	0/258	2/25
9	0/263	0/263	258/263	6/263	0/258	0/6
10	0/263	0/263	256/263	0/263	0/256	undefined
11	0/263	0/263	54/263	0/263	0/54	undefined

indicator is zero; this is not proof of zero population risk. The registered one-sided per-arm upper bound is 1.13%. Coarse-shape, structural-signature, control-interaction, and secondary registered contrasts also have zero observed differences. Descriptive changes in reference-agreeing unanimity also occur on controls and do not rescue the registered endpoint.

E.2 POST-HOC POLICIES ON THE SAVED REPLIES

Per-judge counts and vote histograms had been inspected before this comparison was fixed. It is post-hoc and uses the unchanged replies; it does not replace the registered 0/263 versus 0/263 primary above.

Every policy first requires all eleven original replies to be returned and recognized. This leaves 258 specified and 255 placebo panels from the assigned 263 per arm. For threshold $k = 6, \ldots, 11$, a policy returns CORRECT if at least k votes say CORRECT, INCORRECT if at least k say INCORRECT, and otherwise abstains. All assigned blocks stay in the denominator. A row with no automatic decision has undefined conditional disagreement; it is not zero conditional risk.

At six votes the fixed-denominator contrast is +39.924 points ($[34.2205,\ 46.0076]$); at seven it is +10.646 points ($[7.224,\ 14.449]$). The outcome remains disagreement with the stipulated three-way reference. In the placebo arm, calling `FALSE` correct can agree with a binary "not entailed" reading, so these are not universal objective errors.

The familywise adjustment covers 54 comparisons: three decision channels, six thresholds, and three probe types. Six and seven votes retain Holm values 2.66×10^{-30} and 3.95×10^{-7}. Eight votes has two observed disagreements and is not supported after adjustment; nine through eleven have no observed contrast. At ten and eleven votes the placebo policy makes no decisions. The 54-test correction is part of a post-hoc analysis chosen after histogram inspection; it does not make the analysis confirmatory.

Control coverage and grouping. Both invariant control types have zero observed reference-disagreeing decisions at every threshold. Their coverage moves with the threshold:

Intervals use 10000 paired source-block resamples. The frozen coarse-shape and structural-signature groupings also keep the six- and seven-vote intervals above zero. These are single-implementation source-pattern sensitivities, not independent replications or evidence across task families. All results are conditional on this selected block set and fixed panel.

Table 26: Post-hoc control coverage. S is specified and P placebo. Reference-disagreeing decisions are zero in every cell; zero observed disagreement is not zero population risk.

	Invariant-INCORRECT decisions		Invariant-CORRECT decisions	
k	S	P	S	P
6	263/263	263/263	65/66	60/66
7	263/263	263/263	65/66	60/66
8	263/263	263/263	65/66	60/66
9	263/263	263/263	65/66	60/66
10	263/263	258/263	64/66	60/66
11	151/263	67/263	59/66	46/66

Table 27: Six-vote policies after deterministic reason coding. Binary entailment changes the matching reference; it is not an improvement against an unchanged three-way reference.

	Neutral reference	Reference disagreements		Automatic decisions	
Standard	for candidate FALSE	S	P	S	P
Three-way	INCORRECT	0/263	0/263	223/263	33/263
Binary entailment	CORRECT	0/263	0/263	227/263	251/263

Removing any one recorded vote cannot create six CORRECT votes in a specified panel that had fewer than six. Every placebo panel with at least seven such votes retains at least six after one deletion. Holding original eligibility and all other votes fixed, the bound is therefore zero specified and between 28 and 105 placebo disagreements over 263 blocks. This is arithmetic on saved votes, not a new experiment or a model-deletion effect.

Reason coding and reference standard. A conservative deterministic coder uses only the recorded reason and queried formula. It can identify entailment, refutation, undeterminedness, or partial claims, and abstains on conflicting or unclassified text. These are expressed claims, not hidden thoughts. Under a three-way standard, refuted and undetermined differ; under binary entailment they both map to FALSE. The reference is changed through the same mapping.

E.3 REGISTERED ENDPOINT AND THRESHOLD POLICIES IN ONE DISPLAY

Table 28 keeps the registered unanimity endpoint and the post-hoc threshold policies for the specified and omitted-convention arms in a single display, so that a policy result is never read in place of the registered null that Section 4 reports.

The coder supplies usable states for 16598 of 21703 scheduled jobs (76.5%). Only 181 of 1973 panels have all eleven reasons coded; 1910 meet the original-reply eligibility gate. Opposite-grade pairs are dependent within panels, and matching partial codes do not identify complete beliefs. Under the binary reference, the placebo arm has 22 fully unanimous reference-matching decisions; these are not new wrong-unanimity events against the original three-way reference.

The coder was developed on 198 recorded explanations. A separate 99-example engineering check hid model, arm, and oracle metadata, but its checker had already seen aggregate results; it was neither outcome-blind nor independent human validation. Two adverb-negation bugs were corrected with prior epochs retained. No disagreement count or corresponding test changed, although some coverage did. Independent implementations compare 1782 cell values and 54 error contrasts. The selected replay covers aggregate policy arithmetic and the pure decoder, not original inference, real-response extraction, harmonization, or bootstrap resampling.

Table 28: FOLIO: the registered result first, then post-hoc threshold policies on the same recorded replies. The registered all-eleven unanimity endpoint is $0/263$ in both arms. Decisions are unanimity decisions, not recognized-reply completeness, which is only the eligibility condition. Post-hoc policies abstain below k; at $k = 10$ and $k = 11$ the omitted arm decides nothing, so its zero is abstention, not agreement. These comparisons sit in a family of 54 tests and are convention-relative: with the convention absent, `FALSE` can reasonably mean "not entailed" rather than "refuted." A post-hoc policy result does not replace the registered null. S is specified and O omitted convention.

Rule	Disagreements / 263		Decisions / 263		Multiplicity
	S	O	S	O	
Registered unanimity (all eleven identical)	0	0	54	0	registered, $p = 1.0$
Post-hoc $k = 6$	0	105	258	255	Holm $p = 2.66 \times 10^{-30}$
Post-hoc $k = 7$	0	28	258	102	Holm $p = 3.95 \times 10^{-7}$
Post-hoc $k = 8$	0	2	258	25	Holm $p = 1$, unsupported
Post-hoc $k = 9$	0	0	258	6	neither arm disagrees
Post-hoc $k = 10$	0	0	256	0	O decides nothing
Post-hoc $k = 11$	0	0	54	0	strictest rule; S coverage falls

Table 29: Retained collection settings. Token caps are completion budgets, not guaranteed visible lengths. No model sampling seed was passed.

Protocol	Roster	Cap and collection
Registered paired study	Eleven-model rich roster	1,024; one draw requested, inherited five-draw third retained and disclosed
SQL/regex post-verdict	Eleven-model rich roster	1,024 per turn; one conversation per cell; retries identified separately
Invariant intervention	Twelve-model ladder	1,024; one draw per arm and cell
FOLIO formal test	Fixed eleven-model panel	First returned reply; isolated GPT-5.5 interface exception recorded in Appendix E
Code verification	Separate six-critic roster	Five 512-token critic votes, 1,500-token test generation, 2,000-token reference generation

F Retained protocols, decision definitions, and reference limits

F.1 Rosters, budgets, and parsing

The twelve-model ladder contains Claude Haiku 4.5; Claude Sonnet 4.5, 4.6, and 5; Claude Opus 4.6, 4.7, 4.8, and 5; GPT-5.4 Mini, GPT-5.5, and GPT-5.6; and MAI Code Flash. The eleven-model rich-response roster omits Sonnet 4.5. These are recorded deployment names, not hashes of provider weight snapshots. The code-verification study uses a separate six-critic roster: Claude Opus 4.5, 4.6, and 4.7, plus GPT-5, GPT-5.4, and GPT-5.5.

The primary gateway accepts a temperature argument for interface compatibility but does not forward it to the reasoning-model client. Those runs use provider-managed sampling. The separate verification path passes temperature 1.0. Source-specific completeness and selectors are reported with each study rather than treating protocols as interchangeable.

Judging and probe formats. The SQL prompt supplies schema, question, and candidate, asks whether the candidate is CORRECT for all valid data or BUGGY, and requests one word. The rich collector retains that text and asks for exactly three lines: a 0–100 probability of correctness, a CORRECT/BUGGY verdict, and one short rationale. The parser extracts the verdict, clamps a parsed integer probability to $[0, 100]$, and retains the first rationale line up to 400 characters. Missing fields remain missing.

The post-verdict turn retains the original conversation and asks whether the wording determined the required behavior or left a genuine choice. It requests exactly DETERMINED or OPEN, the adopted reading when open, and one short reason. The regex protocol replaces the SQL framing with Python

3 `re.fullmatch` and otherwise uses the same rich response and turn-two structure. Parsed OPEN is an elicited response category, not a human-validated correctness statement or evidence of prior awareness.

Constraint judging template. The frozen constraint renderer produces the entire `prompt_text`, and the collector passes it directly in one user message. It adds no separate judging wrapper. The two arm headers are:

> **Stated:** You are validating one record against a specification. The specification below is COMPLETE and EXACT: the record is VALID if and only if it satisfies every listed constraint, and INVALID if it violates one or more.
>
> **Withheld:** You are validating one record against a specification. Decide whether the record is acceptable under this specification.

Each header is followed by the same layout:

```
SPECIFICATION (30 constraints):
[30 numbered, arm-specific constraint lines]
RECORD (JSON):
[the shared record]
Answer exactly one word:  VALID or INVALID.
```

The bracketed lines are placeholders, not additional instructions. The header and all constraint lines change wording; the JSON record and executed strict reference stay fixed. The collector normalizes `VALID`/`INVALID` to `CORRECT`/`BUGGY`. This template and direct-message path match the code at the pre-collection `replication-freeze-20260904` revision. Figure 2 shows one actual record excerpt and its retained votes.

F.2 Execution-guard definition

Each critic proposes eight Python assertions based on the specification rather than the candidate; at most ten extracted assertions are retained. Only assertions that pass the trusted clean reference enter plain verification. For a guarded check:

1. Run retained assertions on the audited implementation. If none fails, return CORRECT when at least one test is valid, otherwise NOVALID.
2. For each failing assertion, run a separately generated reference pool on that input, leaving the audited implementation out of its own pool. Ignore unsuccessful executions. A consensus needs a unique plurality with at least two agreeing references.
3. If any resolved failing test has consensus different from the audited output, return BUGGY; an audited crash also differs from a successful consensus.
4. Otherwise, if at least one resolved failing test agrees with the audited output, return CORRECT. If failures exist but none can be resolved, return ABSTAIN.
5. Aggregate critics by OR: any BUGGY gives a panel flag; otherwise any CORRECT gives CORRECT; otherwise abstain.

Plain verification flags any retained test failure. The guard is paired with it on stored implementations, but is not oracle-free and does not resolve every test. Reference validity, abstention, coverage, and lost catches are reported in Appendix D.4.

F.3 Either-reading re-scoring

Withholding an answer-determining clause means a judge can adopt a defensible reading that differs from the designated reference. Alternative executable references were written for the two of six SQL classes that admit a plausible second reading. Scoring a candidate as acceptable when it matches either reference leaves eleven of the nineteen strong reference-disagreeing majorities standing (twelve after relabeling); restricting the licence to the strongest class leaves thirteen. This bounds

the reference-choice explanation but does not turn the remaining cases into a universal objective-error rate.

A judge-level rule that counts a vote wrong only when it is wrong under both readings returns zero, but is uninformative: the two readings assign complementary labels to all 72 items. The same rule would excuse every error in the explicit arm. The judge-level licence is therefore unidentified.

A paired-candidate check presents one candidate for each reading in separate judgments. It finds 89 alternative-coherent judge-pairs in the ambiguous arm and none in explicit or placebo. Individual accept-both behavior rises from 0.2% to 7.2% ($p = 1.8 \times 10^{-4}$), but the panel endpoint moves only $0/35$ to $2/36$ (McNemar $p = 0.50$), and no majority accepts both. The retained claim is failure to recover a designated unstated policy, not a shared wrong answer under every reasonable reading.

G MATCHED WITHHELD-RULE EXTENSION

G.1 FIXED CASES, FIVE MATCHED ARMS

The extension reuses all 160 cases from the policy-slot study in Appendix B: 128 ambiguous and 32 determinate cases across the same eight mechanisms. Case selection is therefore informed by the earlier study, not a new draw of independent mechanisms. The extension's inputs, arms, parser, outcomes and acquisition settings were fixed before its first model call. Every case is kept; none was selected for a previous error or unanimity event.

Four judges – GPT-6 Astra, GPT-5.6, Opus 5 and Opus 5.5 – receive five fresh-context calls per case. S supplies the designated policy and A its opposite, using the earlier policy wording. F supplies generic filler, N announces that the note gives no convention, and R is an exact repeat of N. Both explicit arms are acquired in the same interleaved run as the omission arms, rather than importing old responses as controls. No new paraphrase arm is acquired.

The common header is:

> Evaluate the candidate answer using only the supplied task, input, candidate, and policy note. Judge only those supplied materials.

It is followed by the unchanged output contract: one JSON object with `verdict` and `reason`, a verdict of `CORRECT`, `INCORRECT` or `UNDETERMINED`, and a reason of at most 25 words. The header does not prescribe an `UNDETERMINED` response under omission. The two neutral notes are:

> **N and R:** This policy note states no controlling convention for the present evaluation task.
> **F:** The task and candidate answer are presented below for your evaluation here.

Each note has twelve words. S, A, F, N and R are padded with ASCII spaces to the same character width. Within a case, the header, task, input and candidate are identical across arms; only the note varies. Matching words and characters does not imply matching provider tokens.

Rounding and sorting use the same revised tie instruction in every arm: the expected output follows one fixed tie convention and requires its exact result, not any of several allowed alternatives. This removes a dangling reference to a policy "stated below" while retaining the distinction between a fixed but omitted convention and a set-valued task. The two executable reference functions certify opposite candidate labels on the ambiguous cases and a common label on the determinate controls. The latter are determinate within this two-convention family. No reference output, oracle verdict or policy menu is supplied to a judge.

G.2 ACQUISITION, PARSING AND PLANNED OUTCOMES

The 3200 assigned calls produced 3151 valid replies, 41 parser failures and 8 acquisition failures, over 3207 physical attempts. The completion/reasoning cap is 1,024 tokens. OpenAI uses low reasoning effort; Opus 5 disables thinking; Opus 5.5 uses adaptive low effort. Temperature, top-p and seeds are not sent, and provider defaults are not assumed equal. Jobs are shuffled before collection; each model has at most two calls in flight. Only pre-content connection failures and

specified transient HTTP errors permit one identical retry. Partial HTTP-200 streams, truncation, parser failure and unfavorable verdicts are not rerolled.

The primary parser is the unchanged one-object extractor from Appendix B.2; the earlier exact-format parser is a same-response sensitivity. All calls retain their assigned denominator. A nonvalid reply is neither a vote, an abstention nor a correct judgment. The verdict label describes the *candidate*: `INCORRECT` is the reference-correct judgment when that candidate is incorrect.

The primary comparison is N-versus-S verdict change on ambiguous judge–cases, with the corresponding determinate comparison and N's three-way partition. For valid ambiguous N replies, UNDETERMINED, agreement with S's reference label and agreement with A's reference label exhaust the possibilities. The binary labels are compatible with one reference and incompatible with the other; they do not identify an internal convention or establish universal error. Secondary readouts include the F partition, N/F wording difference, exact N/R repeat stability, explicit-pole correctness, all-arm panel categories and the prespecified model/mechanism breakdowns.

For a pair-change outcome, let c be observed changes, v valid pairs and n assignments. The valid-pair rate is c/v; the assigned observed rate is c/n, with exact allowed-label-completion bounds $[c/n, (c+n-v)/n]$. Ambiguous-minus-determinate bounds subtract the upper control bound from the lower ambiguous bound, and vice versa. These are fixed-assignment sensitivity bounds, not sampling confidence intervals. Mechanism-bootstrap checks resample the eight observed mechanisms 2,000 times; they describe finite-mechanism stability, not a task population.

G.3 Complete results and the absence-notice contrast

The primary changes are $429/496$ valid ambiguous pairs versus $2/122$ valid determinate pairs. The assigned specificity is 82.23 percentage points with bounds $[77.54, 85.35]$. N gives 375 undetermined replies out of 511 valid ambiguous replies. Its remaining 136 binary labels split $68/68$ between the two reference labels; one assignment is nonvalid. F gives 203 undetermined, 149 S-compatible and 157 A-compatible labels, with three nonvalid assignments. This partition concerns all valid replies, not just the changed N/S pairs.

Table 30: A missing-policy notice changes the response without choosing a policy. F is generic filler; N announces that no convention is supplied; R repeats N. U denotes UNDETERMINED; "binary" means four valid equal CORRECT/INCORRECT verdicts. Each mechanism has 64 assigned ambiguous judge–cases and 16 assigned panels, including nonvalid outcomes.

Mechanism	F: U	N: U	F: binary	N: binary
Interval endpoints	2/64	37/64	13/16	0/16
Name matching	38/64	51/64	0/16	0/16
Intersection multiplicity	20/64	48/64	6/16	0/16
Rounding ties	56/64	59/64	0/16	0/16
Timestamp dates	2/64	29/64	14/16	0/16
Graph direction	12/64	34/64	8/16	0/16
Sorting ties	37/64	62/64	0/16	0/16
Empty-group average	36/64	55/64	0/16	0/16
Total	$203/512$	$375/512$	$41/128$	$0/128$

The assigned N-minus-F undetermined increase is 33.59 points, with bounds $[33.01, 33.79]$. Every mechanism's lower bound is positive; the smallest is 3.125 points for rounding ties. The categorical F panels occur in four mechanisms: interval endpoints (13), intersection multiplicity (6), timestamp dates (14) and graph direction (8). All 41 are complete mixed panels under N, not unanimous undetermined panels. The 124 cases complete in both F and N give the same 41-versus-0 comparison. There are 125 complete F panels and 127 complete N panels out of 128 assignments. Allowing every incomplete panel any category gives a 32.03-point decrease with bounds $[31.25, 34.38]$.

Across all ambiguous cases, N has 45 unanimous-undetermined panels, 82 complete mixed panels and one incomplete panel; R has 44, 80 and four, respectively. Neither has a categorical unanimous panel. N/R differ on $31/507$ valid judge–case pairs; N/F differ on $182/508$. One is repeated acquisition and the other changes the note; they are not interchangeable noise estimates. On the

determinate controls, N has 124 valid replies, all reference-correct and none undetermined. R has 126 valid replies, also all reference-correct with none undetermined.

Both explicit S/A labels match their references on $484/512$ assigned ambiguous profiles ($484/487$ valid pairs).

Table 31: Every judge increases undetermined replies under the absence notice. The first three columns use 128 assigned ambiguous judge–cases per model; the last column states its valid-pair denominator.

Judge	F: U	N: U	Both S/A correct	N/R changes
GPT-6 Astra	$67/128$	$111/128$	$126/128$	$14/128$
GPT-5.6	$36/128$	$91/128$	$128/128$	$8/128$
Opus 5	$26/128$	$52/128$	$105/128$	$4/124$
Opus 5.5	$74/128$	$121/128$	$125/128$	$5/127$

A separately labeled exploratory diagnostic conditions on F-categorical panels and finds 37 of the 41 cases with all four judges correct under both explicit poles. It is not the headline denominator or a replacement for the planned N-based categorical witness count, which is zero. The strict-parser sensitivity retains 40 F-categorical panels versus zero N-categorical panels. Under the strict parser, the assigned N/S ambiguity-specific contrast is 76.37 points with completion bounds $[60.74, 85.55]$. The primary-parser mechanism-resampling interval is $[77.34, 87.11]$ points; it describes the eight observed mechanisms, not population coverage.

This is a same-case omission comparison, not discovery of arbitrary hidden policies. Its original single-note result is complemented, not replaced, by the wording controls below. Those controls show why interruption of unanimity cannot be attributed specifically to declaring absence.

G.4 Post-review exploratory vote composition

This analysis was conducted after review of the omission result. It uses only saved F/N/R replies, makes no new calls and does not replace any registered endpoint. A strict binary majority requires more than half the full roster to give the same binary label, with all roster replies valid for these readouts. A three-judge unanimity requires all three retained judges to give the same binary label. Incomplete panels remain a separate category.

Table 32 decomposes the 41 cases selected by their original four-judge binary unanimity under F. Every changed N reply is U, not the opposite binary label. This outcome-conditioned subset is descriptive, not an estimate for unseen tasks. Judge-level changes are likewise subset counts, not a ranking of judge quality or evidence of internal conventions.

Table 32: The 41 originally F-unanimous cases under N, followed by the judges whose responses change. Binary residues retain the original F label; there are no binary-label flips. "Sole U" counts one-U panels in which that judge alone interrupts unanimity.

U votes	Unchanged binary votes	Cases
1	3	15
2	2	16
3	1	10

Judge	Binary-to-U changes	Sole U
GPT-6 Astra	27	6
GPT-5.6	15	0
Opus 5	0	0
Opus 5.5	35	9

The full-roster counts (Table 33) and leave-one-out counts (Table 34) show that a zero four-judge unanimity count does not imply zero three-judge unanimity or zero majority decisions. All per-case

Table 33: Four-judge decision categories on all 128 assigned ambiguous cases. Strict binary majority, U majority and no strict majority partition the complete panels. Binary unanimity is a subset of binary majority, not another disjoint category. Incomplete panels are not successful decisions.

Arm	Complete	Incomplete	Binary unanimous	Strict binary majority	U majority	No strict majority
F	125	3	41	54	39	32
N	127	1	0	16	90	21
R	124	4	0	14	88	22

Table 34: Leave-one-judge-out binary unanimity. Each event count is over all 128 assigned three-judge panels, not only the 41-case subset. Availability is reported separately. These are decision counts, not error rates or deployment gains. Roster removal is exploratory, not a prospective judge-selection rule.

Dropped judge	F unanimous / assigned	N unanimous / assigned	F complete / assigned	N complete / assigned
GPT-6 Astra	43/128	6/128	125/128	127/128
GPT-5.6	43/128	0/128	125/128	127/128
Opus 5	44/128	0/128	127/128	127/128
Opus 5.5	48/128	10/128	126/128	128/128

vote compositions and the R leave-one-out readouts remain in the anonymous replay; no provider-independence or causal attribution follows.

G.5 Prospective wording controls on reused cases

The wording follow-up fixes six arms before acquisition: original filler F, original notice N_0, paraphrases N_1–N_3, and convention mention M. It takes two of four ambiguous cases and the sole determinate case within each mechanism-by-designated-policy-by-candidate-correctness cell using a frozen metadata-only hash ranking. This gives 64 ambiguous and all 32 determinate cases, with no filtering by past verdicts or unanimity. They are reused policy-slot cases, not the disjoint matched-call cohort and not new mechanisms.

The task, candidate, common header and output contract are unchanged. For these cases, F and N_0 prompts exactly reproduce the earlier F and N prompts, but replies are reacquired, not imported or pooled. Table 35 prints the notes with repeated spaces normalized for readability. Actual slots have twelve whitespace-separated words and 99 ASCII characters. The output contract continues to allow U without prescribing it. Matching words and character width does not match model tokens. Semantic equivalence of the notices was a pre-data textual judgment, not an empirical invariance guarantee.

Table 35: All six prospectively fixed note texts. F is generic filler; M mentions convention relevance without declaring absence. N_0–N_3 state that this note does not supply the controlling convention. None supplies the convention, an oracle verdict or an abstention instruction.

Arm	Note text (space-normalized)
F	The task and candidate answer are presented below for your evaluation here.
M	Conventions are relevant to evaluating the task and candidate presented here today.
N_0	This policy note states no controlling convention for the present evaluation task.
N_1	No governing convention is specified in this accompanying note for this task.
N_2	The provided note leaves the task's controlling convention unspecified for this evaluation.
N_3	No task-governing convention has been given in this note for evaluation here.

All four judges use the settings and frozen primary/strict parsers described in Appendix B.2, with a 1,024-token completion/reasoning cap. The six arms are interleaved within outcome-blind shuffled case–judge blocks. Only eligible pre-content/transient acquisition failures permit an identical retry. The 2304 assignments yield 2280 valid outcomes, 23 parser failures and 1 terminal acquisition

failure over 2310 attempts. Missingness is not U. Full saved texts, strict sensitivities, transitions and model/mechanism strata are replayable.

The primary contrasts compare each notice with F and with M on the assigned ambiguous profiles; no success threshold or winning-wording selection was specified. For notice n and control c, observed U counts u_n, u_c and nonvalid counts m_n, m_c on $N = 256$ assignments give

$$\Delta_U = (u_n - u_c)/N, \qquad [\,(u_n - u_c - m_c)/N,\ (u_n + m_n - u_c)/N\,].$$

Observed valid labels remain fixed and each unknown can take any allowed label. These exact completion bounds are sensitivity ranges for the assigned records, not population confidence intervals.

Table 2 reports every arm's U, reference-correct and valid counts with assigned denominators; Table 3 gives all eight contrasts and their bounds. Complete ambiguous rosters, in F/M/N_0/N_1/N_2/N_3 order, number $60, 61, 62, 63, 64, 64$ out of 64. F has 19 binary-unanimous panels, with completion-count bounds $[19, 20]$; M and every notice have none, bounded by $[0, 0]$. Their missing cells cannot restore binary unanimity while holding observed votes fixed.

All fixed wordings are reported, rather than selecting a favorable one after observing the outcomes. Section 4 retains the convention-mention control, determinate costs and model-level non-separation beside the aggregate contrasts; those limits also govern interpretation of this appendix.

H ANONYMOUS SUPPLEMENTAL REPLAY AND PROVENANCE

The accompanying `anonymous_offline_replay_v2.zip` has 50 members. It covers the policy-slot study (1,920 assigned cells), disjoint matched-call study (1,536), withheld-rule extension (3,200) and wording follow-up (2,304): 8,960 assignments in total. The post-review exploratory vote analysis reuses the withholding records and adds no assignments. The older historical, human-coding, execution and FOLIO studies are outside this ZIP.

Project-level provenance also retains generators, fixed seeds, original draw records and frozen analysis for older studies. Excluded calibration and breadth studies are preserved in a separate unsubmitted TeX archive, not this paper's evidence package. Provider deployment names do not identify reproducible random streams or weight snapshots. Model and human codes remain distinct, and original human responses are preserved outside the ZIP.

After extraction, run the following from its `anonymous_replay` directory with Python 3.10 or later; only the standard library is required:

```
python -I -S -B replay.py
```

The command writes `replay_output`, validates payload hashes, tests parsers, recomputes executable references, reparses terminal outcomes and compares regenerated tables with the closed records. It has been replayed successfully in a clean local directory. Process-local guards reject network and child-process operations; there are no new model calls.

Exact decoded prompts and saved response texts accompany every available attempt. Failure statuses and all assigned denominators are retained: absent text is not an empty answer, U or a valid vote. Returned text is not a provider token stream; documentary stream hashes are distinguished from hashes of the exported text. The ZIP includes the offline scientific functions and adapters, not private acquisition clients. Replaying observations does not regenerate provider sampling, weights, execution tests or human annotations.

Table 36: Prospective local freeze records for the controlled studies. Times are shown to seconds in UTC; full precision and source-hash mappings are in `PROVENANCE.json`. Each SHA-256 is printed on two consecutive lines that must be concatenated. These are local provenance records, not external registrations or independently certified timestamps. The exploratory vote analysis has no prospective freeze claim.

Study	Local seal time and full lock SHA-256

Policy-slot and matched-call studies (M45)	2026-09-25 14:08:37 UTC
	`d1a55b1fdb8f74c83776e699f47c7045`
	`5f342484771f81abb688eebd0483661b`

Withheld-rule extension (M54)	2026-09-26 01:57:29 UTC
	`48c7f0afd2f833fc558b68a4c16a9e29`
	`9cb430c0f62ffbacc7aa090d5216eca3`

Notice-wording follow-up (M59)	2026-09-26 08:18:32 UTC
	`08039b3b62e9e83aaaf872d4c5dffd3c`
	`7d44d9f6ecdeccfae0d4820ef904d122`

Archive SHA-256 (concatenate the two lines):

```
01ff39aa55d59d71a15ffb3d02e7b559
fb54a46f454e880cb1db8167568cdf73
```

This identifies the accompanying ZIP, not a public website or upload. No external preregistration or license grant is asserted. The supplement excludes credentials, private transport, Git history and human records. Its scientific content can still support inference about the study; anonymity checks are not a guarantee against such inference.