A quoted answer overrides the rule that designates it: what an instruction-compliance score measures depends on how the rule names its answer

Anonymous authors
Paper under double-blind review

Abstract

Conflict-only instruction-compliance scores ask whether a verifier endorses the rule-designated answer when the facts favour another. Their contrast, $J_X = p_{10} - p_{01} = R_C - R_T$, subtracts correctness sensitivity from compliance sensitivity; without aligned items, it cannot identify the two separately. On four fixed instruction-tuned checkpoints, holding label indirection fixed and printing the permitted answer changes conflict-cell error by $Q_{\text{quote}} = 0.153$ $[0.127, 0.180]$, with substantial variation across checkpoints.

We then separate the rule's label-based designation from the answer it quotes. In a prospectively registered form where the two disagree, and with the factual question removed and both options replaced by invented words, accuracy against the label binding is 0.103, versus 0.955 when they agree. The quoted option is endorsed at 0.806, versus 0.012 for the label-designated one. The label binding remains usable without a quotation (balanced accuracy 0.922); thus factual-answer recall cannot explain this override of the label-based designation in these constructed rules.

For property-based rules, balanced accuracy without a factual question is 0.526 $[0.518, 0.535]$: below the registered competence floor, but not at chance under its registered criterion. The verdict is indeterminate, so their conflict-cell cost cannot be isolated as a rule-form effect. An independent equal-budget comparison also did not establish a selection advantage for the four-cell audit. These empirical findings concern this fixed small-model roster and two-option tasks; the identification problem applies to the score's definition.

1 Introduction

A verifier is often scored on instruction–prior conflict alone: the rule designates one answer, the facts support another, and we record how often the model endorses the compliant candidate (Longpre et al., 2021; Wu et al., 2024; Kortukov et al., 2024; Zhang et al., 2025; Xu et al., 2024). Our claims have three levels. First, the conflict-only score cannot separate sensitivity to the rule from sensitivity to factual correctness. Second, how the rule names its answer changes that score, though the size and direction vary by checkpoint. Third, when a quotation is experimentally put in conflict with the rule's label-based designation, these checkpoints follow the quoted answer even with the factual question removed. We also report where a property-based rule cannot support the same interpretation.

One number standing in for two. Let C mark compliance with the designated answer and T baseline correctness, and let p_{ct} be the endorsement rate in cell (c, t). The conflict contrast is $J_X = p_{10} - p_{01}$. Writing R_C for the endorsement change associated with compliance at fixed correctness and R_T for the change associated with correctness at fixed compliance, $J_X = R_C - R_T$. A verifier that ignores the rule and one that applies it strongly while tracking correctness slightly more produce the same J_X. The two aligned cells are what separate them, and a conflict-only design never collects them. This is short and algebraic, and we treat it as the specification of the instrument rather than as a result.

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.

What the compliance contrast partly reads. In the standard construction the rule prints the compliant answer — your answer must be "Madrid". A model that endorses whichever string the rule quotes scores $R_C = 1$ and $R_T = 0$, indistinguishable from one that applies the rule. We close that shortcut with rules that designate an answer by an arbitrary label and print neither candidate. A prospectively registered form then makes the label-based designation and the quotation disagree. With the factual question removed, every checkpoint endorses the quoted option over the label-designated one. Here "override" names that observed preference under a deliberately inconsistent rule, scored against its label binding.

Half the roster cannot read a prohibitive rule. Strip the world-knowledge question away entirely, leave only two labelled options and a rule, and ask the same yes/no question. On the permissive form (only the answer labelled M is permitted) all four models are above the midpoint between chance and perfect. On the prohibitive form (the answer labelled K is forbidden) Llama-3.2-3B sits at 0.494 — chance — and never once says yes when the rule permits the candidate; Qwen3-4B reaches only 0.613. Compliance scores collected on forbid-style rules are therefore partly a measurement of whether the rule was parsed at all, and we report every compliance number for a model beside its result on this arm.

What we did. We rebuilt the four-cell design on 160 new base questions in 40 construction groups sharing no fact with an earlier unpublished study of ours, under three rule forms and three verifier wordings, pre-registering endpoints, weights, clustering and two joint practical thresholds before the first generation existed, with a coverage simulation at the frozen design. We then tested whether filling in the aligned cells buys a better verifier than a conflict-only audit at equal budget, on 400 independent rule instances whose labels come from a deterministic predicate rather than from another model.

Contributions. (i) A single token-matched manipulation yields all four cells by flipping which answer the rule designates while holding the question, candidates and option order fixed. The algebra is signal detection theory on a 2×2 (Green & Swets, 1966; Macmillan & Creelman, 2005), which we do not claim as new; the design separates the two contrasts, and validating R_C on these same cells would only change units ($\mathrm{BA}_4 = \frac{1}{2} + \frac{1}{2}R_C$). (ii) Pre-registered comparisons separate quotation from label indirection and probe explicitness and matchability. Further discriminating forms were filed with competing accounts and predicted signs before generation; the registered aligned-cell specificity control could not have failed, and we report that limitation. (iii) The relation-only pseudoword intervention separates quotation from label designation without a factual answer. This is the headline empirical result, scoped to these checkpoints and constructed rules. (iv) An independent equal-budget selection experiment returns a null: on this roster the aligned half of the audit could not have won. Thus the four-cell design is motivated here, while a practical selection benefit remains unestablished (Table 1; Section 4.4).

2 Related Work

Shortcut learning and annotation artefacts: same instrument, different object. Benchmarks are routinely solvable by cues that stand in for the task — hypothesis-only baselines recover much of NLI accuracy (Gururangan et al., 2018; Poliak et al., 2018), lexical overlap substitutes for inference (McCoy et al., 2019), a single unigram carries an argument comprehension benchmark (Niven & Kao, 2019), and Geirhos et al. (2020) give the general account. Our diagnostic — change one thing, hold the item fixed, re-measure — is that tradition's, as is the stimulus discipline (Kaushik et al., 2020; Gardner et al., 2020); we claim neither. What differs is where the cue lives. There it is a statistical regularity of the data, fixed by collecting data differently; here it is in the rule text of the measuring instrument, which writes the answer it is about to ask for — a property of the protocol, present at any benchmark size, and removable by editing one sentence.

Judging instruction compliance is the setting; its known failure modes are other ones. Model judges are the standard instrument (Zheng et al., 2023; Zeng et al., 2024), with verifiable-instruction suites scoring constraints programmatically (Zhou et al., 2023). Their

documented pathologies are about the judge — position and verbosity effects (Wang et al., 2024); ours is about the item: with the judge fixed, how the rule is worded moves the score, in a direction its reader cannot see. Negation is the sharpest case, and it is known to be hard for language models (Kassner & Schütze, 2020; Ettinger, 2020; Truong et al., 2023); §4.3 reports that half our roster cannot apply a prohibitive rule with no conflicting fact present at all. Zeng et al. (2024) is the nearest methodological neighbour — adversarial judging items whose surface cues mislead the judge — but their cues are properties of the candidate output and ours is in the rule.

Prompt sensitivity: the same magnitude, a different kind of cue. Semantically inert formatting swings scores and reorders leaderboards (Sclar et al., 2024; Mizrahi et al., 2024; Alzahrani et al., 2024); our effect is of that magnitude and not of that kind, being content-bearing and directional — it decides which answer is endorsed rather than adding variance. Instruction-hierarchy work asks what a model does when instructions conflict (Wallace et al., 2024; Zverev et al., 2025; Zhang et al., 2025); R1Qx is a conflict within one instruction, and the quotation wins.

Instruction–prior conflict is established; what a score over it identifies is not. Models override or fail to override a baseline answer when an instruction demands a different one, at rates depending on model, task and training recipe (Wei et al., 2023; Goyal et al., 2025; Wang, 2026), and the conflict has its own construction methods and survey (Longpre et al., 2021; Xie et al., 2024; Xu et al., 2024). That literature measures behaviour under conflict; our question is one level out — given a score collected under conflict, what about the verifier does it pin down? — and the answer is an identity, $J_X = R_C - R_T$, which is why the aligned condition is not optional.

Adjacent literatures. Self-critique fails on hard tasks (Huang et al., 2024; Tyen et al., 2024; Kamoi et al., 2024; Stechly et al., 2025; Binder et al., 2025), response habits move scores (Zheng et al., 2024; Braun, 2025), and probes recover what the prompt makes salient (Farquhar et al., 2023); none concerns what an endorsement rate identifies when the requirement is in the context window, and Appendix S10 separates each. Nothing here rests on mechanistic localisation and no causal language is used.

3 Design and estimands

A verifier is asked a question of the form: here is a requirement, here is a candidate answer, does the candidate satisfy the requirement? When the requirement and the facts agree, answering well is easy and uninformative. The interesting case is conflict: the requirement designates an answer that is not the baseline-correct one. Scores are therefore usually collected on conflict items alone.

3.1 Two labels, four cells

Each stimulus carries two binary labels: C, whether the candidate complies with the requirement, and T, whether it is the baseline-correct answer. Let $p_{ct} = \Pr(\text{yes} \mid C = c, T = t)$. Holding the question, both candidate strings, their correctness labels and every other token fixed and changing only which answer the requirement designates flips C for both candidates and leaves T alone, so one manipulation yields all four cells from one base question. Write $J_A = p_{11} - p_{00}$, $J_X = p_{10} - p_{01}$, and

$$R_C = \tfrac{1}{2}(J_A + J_X), \qquad R_T = \tfrac{1}{2}(J_A - J_X), \qquad \text{so that} \qquad J_X = R_C - R_T.$$

R_C is the change in endorsement associated with compliance at fixed correctness and R_T the change associated with correctness at fixed compliance; both are response-frequency contrasts, not capacities and not causal effects of truth. J_X is one number and $R_C - R_T$ is a difference of two, so a verifier that ignores the requirement and one that follows it strongly while following correctness slightly more produce the same J_X, and no transform of a conflict-only score supplies the aligned rates it never observed (Figure 1).

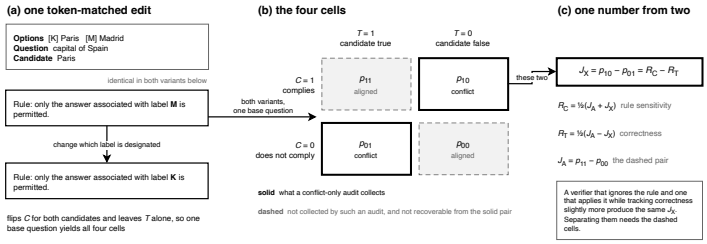

Figure 1: A conflict-only score reports the two solid cells, and the contrast it forms from them is a difference of two quantities it cannot separate. Flipping which label the rule designates, with everything else held fixed, changes C for both candidates and leaves T alone, so one base question yields all four cells. Drawn, not measured: no quantity here comes from a run.

None of this algebra is new, and we do not claim it. Read endorsement under C=1 as a hit and under C=0 as a false alarm and the four cells are two hit rates and two false-alarm rates at the two levels of T, with $\mathrm{BA}_4 = \frac{1}{2} + \frac{1}{2}R_C$ the familiar $\frac{1}{2}(H + 1 - F)$: signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005) on a 2×2, whose standard main-effect-versus-interaction statement is "one contrast cannot recover two". Two consequences are ours and are design points rather than identities: a single token-matched manipulation yields all four cells, and "R_C predicts four-cell accuracy" is therefore a change of units rather than a validation, so any downstream test must run on independent items with independently determined labels, which is what §4.4 does.

3.2 The shortcut a quoting rule leaves open

In the usual construction the requirement prints the compliant answer verbatim — your answer must be "Madrid". This is not our stipulation about the field: of the 24 verifiable-instruction types released with IFEval, 10 state the requirement by printing the literal string a compliant response must contain and 1 by printing one it must not, against 13 that name a property — a count, a format, a language — with no string in the rule to match. The canonical instruction–prior conflict task prints it too: Memo Trap (McKenzie et al., 2023) asks for a quote that ends in the word "heavy". We audited 13 suites from their own repositories, papers and dataset cards under a coding rule fixed before counting (Appendix S11, Figure 2). It is a statement about rule-writing conventions and nothing more, and the population it can be a statement about is itself a finding: of the 13 suites, 5 decide compliance programmatically and 7 by a model judge — but only 2 of the judged ones publish a per-item rule template at all, the rest stating each requirement in free-form text. So the enumerable population is the one that cannot exhibit the effect, and the population that can is not published in a form anyone can audit. We ran no model on any of them and claim no published score is inflated. A verifier that endorses whichever string the requirement quotes produces $R_C = 1$ and $R_T = 0$ and is indistinguishable from one that applies the rule: R_C measured this way is a contrast, not a capacity, and that decides whether a compliance score is about rule application at all.

We close that specific shortcut with rules that designate an answer through a label and never print either candidate string — panel (a) of Figure 1 shows the stimulus in full: a two-option block, a rule permitting the answer associated with one label, and one candidate answer.

We use three rule forms. R0 quotes the designated answer and is the bridge to the earlier construction. R1 permits the designated label. R2 forbids the other label, so the rule names the label of the non-compliant answer; the two-option sentence is mandatory under R2, since without it forbidding one option does not designate the other.

This closes one shortcut — a candidate string appearing in the rule — and not every shortcut, and it is strictly harder, since resolving a label is an extra step. Section 4.3

measures that step on its own, because a model that cannot resolve a label produces errors here that say nothing about instruction–prior conflict.

Crossing the two things R0 → R1 changes at once. That sentence names a confound we did not originally measure. Moving from R0 to R1 removes the quoted string and adds the indirection step, so a change in J_X between them cannot be assigned to either. We therefore add the missing cell, in which the rule is indirect and quotes:

R1Q: Only the answer associated with label M ("Madrid") is permitted.

R1Q is R1 with one insertion and nothing else changed, which crosses the two factors: R0 prints the string without the indirection, R1 has the indirection without the string, and R1Q has both (Figure 2 places every form of this paper on the same axis). So $Q_{\text{quote}} = L_X(\text{R1}) - L_X(\text{R1Q})$ holds the indirection fixed and varies the string, and $Q_{\text{ind}} = L_X(\text{R1Q}) - L_X(\text{R0})$ does the reverse. Rule form is crossed with the base question, so both are within-item contrasts: letters, option order and both candidate strings are identical on either side of each difference.

Two further forms guard the reading. R1QQ prints both strings, separating string presence from binding to the permitted label: under mere presence p_{01} rises, under binding it stays near zero. R1-long is a length control, R1 plus a sentence restating the two-option requirement every form already carries, matched to R1QQ's mean body length — a quote effect that is really a length effect must appear in it too. Forms, estimands and thresholds were registered before any response to them existed (Appendix S1).

3.3 Endpoint, and the sampling unit

Every generation falls in exactly one of yes, no, unreadable; transport failures are recorded separately and never silently converted. The primary endpoint is operational, $e_{ct} = \Pr(\text{no correct readable verification label} \mid C = c, T = t)$, with $L_A = \frac{1}{2}(e_{11} + e_{00})$ and $L_X = \frac{1}{2}(e_{10} + e_{01})$; it imputes nothing about a latent yes or no, and where unreadable occurs we give identification bounds rather than a fill (Appendix S16). Base questions are authored in swap groups of four from one semantic class, so the group, not the item, is the independent sampling unit and one resample moves every row of a group together; roster and template blocks are fixed design factors, so intervals cover new questions and construction groups, not a population of models or prompts (§S21).

4 Results

4.1 What was run

Four instruction-tuned checkpoints, 160 new base questions in 40 construction groups, three rule forms and three verifier wordings: 23040 verification calls, greedy (temperature 0, one per cell, at most 24 new tokens). Every call returned a readable yes or no (0 transport failures, 0 unreadable), so no rate here is a decoding sample and the identification bounds bind nothing. No question or answer is shared with the earlier unpublished study of ours that motivates C1 (withheld for anonymity), a disjointness asserted by test, and a pilot of 24 questions is permanently barred from the set. Ten intervals carry a pre-registered decision and the rest are descriptive (Appendix S24); Appendix S13 lists every departure from the filed plan, including one registered arm we did not run and why.

4.2 Under a quoting rule, models do not reject the string the rule printed

Let p_{01} be the rate of endorsing the candidate the rule did not designate. Under the quoting rule it is exactly zero in 11 of 12 model × wording cells — no endorsement in 160 opportunities — with a mean of 5.0×10^{-4} and a maximum of 0.006. A rule that prints its answer is satisfied by matching that string, and matching is easy. Remove the quote and that disappears: mean p_{01} rises to 0.199 under the permissive label rule and 0.486 under the prohibitive one, with exactly-zero cells falling to 2 and 0.

The conflict contrast changes sign. J_X is positive in 11 of 12 cells under the quoting rule (mean 0.207), and in only 3 and 3 under the two label rules (means -0.067 and -0.367; per-cell and per-wording in Appendix S2). The decomposition says where that comes from: mean R_C — the component usually read as instruction responsiveness — falls from 0.597 to 0.458 and 0.248, while R_T moves from 0.390 to 0.524 and 0.615. The apparent responsiveness was substantially the quote.

It is not simply that the task got harder (C3). The aligned cells were registered as the specificity control; all three pre-declared conditions hold, but at $J_A = 0.988$ they were at ceiling and could not have failed, so we report them as a failed control rather than as the control (Appendix S12).

The quote and the indirection step, crossed directly (E1). The aligned cells are an indirect control: they say a generic difficulty increase would have raised error in both halves, not which of the two changes carries the sign flip, so we added the missing cell and crossed the factors (Section 3.2). Holding the indirection fixed and varying only whether the permitted answer is printed gives $Q_{\text{quote}} = 0.153$ $[0.127, 0.180]$, above the 0.05 threshold registered in advance and excluding zero on 3 of four models individually; holding the string fixed and varying only the indirection gives $Q_{\text{ind}} = -0.055$ $[-0.066, -0.042]$, so the indirection step does not make the conflict cells harder and with the string present makes them slightly easier. The sign change is carried by the printed string. Registered on the bridge wording and run on the two paired wordings as well, Q_{quote} is 0.188 and 0.185 there, positive and excluding zero on 3 of four in each, with Llama-3.2-3B resolved against it on both (Appendix S13, D8).

Property rules limit what the conflict score can establish (addenda 10 and 11). Every form above designates one of the two listed options; the largest class of published verifiable instructions does not, deciding compliance by a property of the response with no string to match and no options block to resolve (Appendix S11, Figure 2). Two such forms, registered before they were run, give $Q_{\text{prop}} = 0.449$ against a rule that prints its answer, on 4 of four checkpoints (Table 1). We do not read that as a rule-form cost. The control registered to license such a reading — the same forms with no question and no true answer — returns balanced accuracy 0.526 against these properties, above addendum 11's filed floor on 0 of four, while the label forms on those same stimuli run 0.922 to 0.987: these checkpoints apply a label rule almost perfectly with nothing competing, but perform near chance on a property. The property control meets neither registered competence nor chance criterion. We therefore report its verdict as indeterminate, withdraw the rule-form reading of Q_{prop}, and cannot interpret the conflict-cell score as a clean measure of rule-form sensitivity on these stimuli (Appendix S15; Appendix S13, D15). This arm also does not settle the self-containedness reading of the decomposition.

The length control moves the other way. R1 plus an inert sentence matched to R1QQ's length gives $Q_{\text{len}} = -0.038$: added length raises conflict-cell error while the quote lowers it, so the registered invalidation — which fires only if the two move together — does not fire, and length works against the quote effect throughout.

What the printed string does, tested prospectively three times. Q_{bind}, the registered contrast separating string presence from binding, excluded zero in the direction a pure compare-and-reject account forbids (0.113). Rather than reinterpret after the fact we filed three further forms with their predicted signs before any existed. The first, R1Qneg, quotes only the forbidden answer: the loose reading endorse whatever string the rule prints predicts p_{01} rises and it falls, 0.158 on 0 of four in the predicted direction, while L_X improves on R1 by 0.224 (Appendix S14).

R1Qx makes the two sources of designation disagree: the rule binds the permitted label as always, and the string it quotes in that slot is the other candidate's answer (Figure 4). We use "override" to mean that endorsement follows the quotation instead of this label-based designation within the deliberately inconsistent rule. Scoring is against the label binding, fixed in the pre-registration before the run. If the label governs, endorsement of

Table 1: Conflict-cell error L_X by rule form, roster mean over the four checkpoints: the numbers behind Figure 3, as named estimands, so a claim about the ordering can be checked here rather than recomputed from the per-cell tables. † scored against the label binding, not interchangeable with the rest; ‡ a property rule, a different rule class (withdrawal above).

form	L_X	form	L_X	form	L_X
R1QQ	0.287	R0	0.456	R1pos	0.726
R1Qneg	0.330	R1	0.554	R1Qsep†	0.870
R1Q	0.401	R1-long	0.592	RpropV‡	0.905
R1Qdesc	0.439	R1Qdescx	0.712	RpropN‡	0.939
				R1Qx†	0.955

the designated candidate stays positive and near its R1Q value of 0.598. It does not merely fall; it inverts: $X_{\text{desig}} = -0.452$ $[-0.469, -0.434]$, negative on 4 of four checkpoints and on 0 in the other direction, with operational error 0.955 against the binding and no unreadable reply from any checkpoint: the models are not confused by the inconsistency, they resolve it.

They resolve it by substituting the quoted string for the designation: the structure R1Q shows for the designated candidate reappears under R1Qx for the quoted one, 0.497 against 0.045 (Appendix S22 gives the R1Q pair and this interval). All four cells are needed, since in the conflict condition the quoted candidate is also the baseline-correct one. The aligned cells separate them: p_{11}, where the candidate is true and designated and only the quoted string points elsewhere, falls from 0.997 to 0.083, which no prior account predicts, while p_{00} stays at 0.077 — the string is necessary and not sufficient.

And it is not the parenthesis, and here is what the main-bank cells can and cannot show. R1Qsep moves the contradicting string out of the label's apposition into its own sentence; the override weakens and survives, $X_{\text{desig}} = -0.356$ on 4 of four, with the apposition worth 0.095 of it (Appendix S26). Pooled over four cells, though, X_{desig} hides a split we report rather than average away: in the conflict half the quoted string is the baseline-correct answer, so endorsing it and answering the question are the same act, while in the aligned half the models endorse neither candidate. What the main bank demonstrably shows is that a contradicting quotation destroys endorsement of the label-designated answer. The relation-only pseudoword arm settles the stronger reading, since no true answer is there for the quoted option to coincide with: the quoted option is endorsed at 0.806 against 0.012 for the designated one, the binding it overrides is operative on those very stimuli (balanced accuracy 0.922 under a label rule with nothing printed), and every checkpoint is below chance on both variants as addendum 6 filed (Appendix S28, Figure 4).

It is the string, not the dereference. One account fits the forms above without mentioning strings: these models resolve an arbitrary letter poorly, and any form computable without that resolution scores better. The property rule looked like a refutation of it and is not one, for the reason just given. Two results weigh against it: designation by position, registered with the prediction that account makes, gives $Q_{\text{pos}} = -0.172$ on 0 of four in the predicted direction, and gating Q_{quote} to items a model demonstrably resolves leaves it unchanged (Appendix S17; post hoc).

How much is matchability, and how much is the extension being written out. Every form so far confounds them. R1Qdesc separates them by identifying the designated answer with a definite description from the frozen bank — extension written out, nothing to match — and recovers most of the effect: $Q_{\text{desc}} = 0.115$ on 3 of four checkpoints, against $Q_{\text{quote}} = 0.153$ on the same items, the difference -0.038 also excluding zero. The per-model counts do not treat the two alike, screened or unscreened: explicitness survives at Q_{quote}'s own count and matchability does not, and since applying the screen to one contrast but not its comparator would be a standard chosen per number, Appendix S23 reports every count both ways. A third form built to break the tie, R1Qdescx, is uninformative as constructed and we draw nothing from it (Appendix S13, D13).

What the five forms say together. Error falls as more of the rule's extension is written out (4 of four checkpoints at the endpoints and 3 for the full order, over the five crossed forms of Table 1; Appendix S2), and where a string is printed the candidate is compared against it in the polarity of its slot — endorsed from the permitted slot even against an explicit binding, refused from the forbidden one. The binding is the fallback when no string is printed, and §4.3 shows how poorly half the roster manages it.

The registered per-model criterion turns on a checkpoint that barely responds at W0. Q_{quote} excludes zero on 3 of four as filed, but Qwen3-8B's mean conflict-cell endorsement never exceeds 0.050 under any of the five forms the screen covers (it does respond to R1Qx, which is that form's point and not this screen's); screening at 0.10 leaves 3 checkpoints, of which 2 exclude zero. The criterion fails under that post hoc screen and we report both counts.

With no prior available (E2). On the relation-only branch, printing the permitted string still helps (0.025 in balanced accuracy, and 0.033 with invented strings, so not recognition of a familiar answer), but the lower bound falls short of the 0.05 margin we registered, so by our own rule E2 is not established. R1Q is at or near ceiling there, so this is headroom as much as attenuation and we do not compare its size with the main bank's.

It is not uniform, and we do not claim it is. Q_X is positive on three checkpoints and resolved the other way on Llama-3.2-3B; the pre-declared rule requires three of four, so the claim names those three (Appendix S8).

4.3 Two pre-registered negatives, and reading a rule with no fact to conflict with

Naming the decision criterion, against a wording matched on length, structure and emphasis rather than against our earlier study's own original question, does not selectively damage the aligned condition ($D = 0.013$, the registered joint rule unmet at a power of 0.955) — a resolved negative (Appendix S5). The second is E2 above. With the question stripped away entirely, all four models clear the midpoint between chance and perfect on the permissive form, while on the prohibitive form Llama-3.2-3B is at chance, 0.494, never endorsing a candidate such a rule permits — which an always-no responder also produces, and this arm does not separate the two (Appendix S19).

4.4 Does filling in the aligned cells buy a better verifier? (C2)

Four audit policies at equal budget picked among twelve configurations on 400 independent rule instances labelled by a deterministic predicate, with selections, scores and the hash of every visible row locked before any transfer output was scored. The pre-specified comparison, the crossed audit against the full hit/false-alarm baseline, is $G = -0.007$ $[-0.019, 0.004]$: it does not exclude zero and favours the baseline. No advantage to the crossed audit is established, and of the four selected configurations the one the scalar conflict audit chose has the lowest test error. Both four-cell policies score by the worst cell, so the aligned half can change a selection only when aligned error rivals conflict error, and here it never comes close: that half is worth its cost when a verifier can fail on aligned items, and on this roster it cannot (Appendix S18 gives the policies, the slices and the per-configuration scores).

5 Limitations, and what we conclude

Scope, and scale runs against us where we can see it. Four checkpoints at 3–8B and one bank of two-option world-knowledge verifications; models, templates and domain are design factors, not samples, and with one checkpoint reversing sign a roster mean is a dependency report. Q_{quote} is smaller on Qwen3-8B than on Qwen3-4B, the only within-family comparison here, and the registered arm above 8B was not run (Appendix S13). Nothing establishes that what remains once the string is removed is rule application in any deeper sense, and the registered difficulty control could not have failed at $J_A = 0.988$. We conclude that most of what a conflict-only score measures here is how the rule was written.

References

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 13787–13805, 2024. URL https://aclanthology.org/2024.acl-long.744/.

Binder et al. Looking inward: Language models can learn about themselves by introspection. In International Conference on Learning Representations (ICLR), 2025. URL https://arxiv.org/abs/2410.13787.

Daniel Braun. Acquiescence bias in large language models. In Findings of the Association for Computational Linguistics: EMNLP, 2025. URL https://arxiv.org/abs/2509.08480.

Allyson Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics, 8:34–48, 2020. URL https://aclanthology.org/2020.tacl-1.3/.

Farquhar et al. Challenges with unsupervised llm knowledge discovery. arXiv preprint, 2023. URL https://arxiv.org/abs/2312.10029.

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. URL https://aclanthology.org/2020.findings-emnlp.117/.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11):665–673, 2020. URL https://arxiv.org/abs/2004.07780.

Sachin Goyal, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. Context-parametric inversion: Why instruction finetuning can worsen context reliance. In International Conference on Learning Representations (ICLR), 2025. URL https://arxiv.org/abs/2410.10796.

David M. Green and John A. Swets. Signal Detection Theory and Psychophysics. Wiley, New York, 1966.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (NAACL-HLT), pp. 107–112, 2018. URL https://aclanthology.org/N18-2017/.

Huang et al. Large language models cannot self-correct reasoning yet. In International Conference on Learning Representations (ICLR), 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Kamoi et al. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. Transactions of the Association for Computational Linguistics, 12:1417–1440, 2024. URL https://arxiv.org/abs/2406.01297.

Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 7811–7818, 2020. URL https://aclanthology.org/2020.acl-main.698/.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In International Conference on Learning Representations (ICLR), 2020. URL https://arxiv.org/abs/1909.12434.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. Studying large language model behaviors under context-memory conflicts with real documents. In First Conference on Language Modeling (COLM), 2024. URL https://openreview.net/forum?id=xm8zYRfrqE.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021. URL https://aclanthology.org/2021.emnlp-main.565/.

Neil A. Macmillan and C. Douglas Creelman. Detection Theory: A User's Guide. Lawrence Erlbaum Associates, Mahwah, NJ, 2nd edition, 2005.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 3428–3448, 2019. URL https://aclanthology.org/P19-1334/.

McKenzie et al. Inverse scaling: When bigger isn't better. Transactions on Machine Learning Research, 2023. URL https://arxiv.org/abs/2306.09479.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. Transactions of the Association for Computational Linguistics, 12:933–949, 2024. URL https://aclanthology.org/2024.tacl-1.52/.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4658–4664, 2019. URL https://aclanthology.org/P19-1459/.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM 2018), pp. 180–191, 2018. URL https://aclanthology.org/S18-2023/.

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), 2024. URL https://arxiv.org/abs/2310.11324.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations of large language models on reasoning and planning tasks. In International Conference on Learning Representations (ICLR), 2025. URL https://arxiv.org/abs/2402.08115.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: an analysis of language models on negation benchmarks. In Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), pp. 101–114, 2023. URL https://aclanthology.org/2023.starsem-1.10/.

Tyen et al. LLMs cannot find reasoning errors, but can correct them given the error location. In Findings of the Association for Computational Linguistics: ACL, 2024. URL https://arxiv.org/abs/2311.08516.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. arXiv preprint, 2024. URL https://arxiv.org/abs/2404.13208.

Han-yu Wang. Persistent priors, preserved targets: A Stroop-style paradigm for lexical override. arXiv preprint arXiv:2606.07555, 2026. URL https://arxiv.org/abs/2606.07555.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 9440–9450, 2024. URL https://aclanthology.org/2024.acl-long.511/.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. arXiv preprint, 2023. URL https://arxiv.org/abs/2303.03846.

Kevin Wu, Eric Wu, and James Zou. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3aa291abc426d7a29fb08418c1244177-Abstract-Datasets_and_Benchmarks_Track.html.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In International Conference on Learning Representations (ICLR), 2024. URL https://arxiv.org/abs/2305.13300.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8541–8565, 2024. URL https://aclanthology.org/2024.emnlp-main.486/.

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), 2024. URL https://arxiv.org/abs/2310.07641.

Zhihan Zhang, Shiyang Li, Zixuan Zhang, Xin Liu, Haoming Jiang, Xianfeng Tang, Yifan Gao, Zheng Li, Haodong Wang, Zhaoxuan Tan, Yichuan Li, Qingyu Yin, Bing Yin, and Meng Jiang. IHEval: Evaluating language models on following the instruction hierarchy. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), pp. 8374–8398, 2025. URL https://aclanthology.org/2025.naacl-long.425/.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In International Conference on Learning Representations (ICLR), 2024. URL https://arxiv.org/abs/2309.03882.

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 (NeurIPS), Datasets and Benchmarks Track, 2023. URL https://arxiv.org/abs/2306.05685.

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, 2023. URL https://arxiv.org/abs/2311.07911.

Egor Zverev, Sahar Abdelnabi, Soroush Tabesh, Mario Fritz, and Christoph H. Lampert. Can LLMs separate instructions from data? and what do we even mean by that? In International Conference on Learning Representations (ICLR), 2025. URL https://arxiv.org/abs/2403.06833.

AI Usage Disclosure

Large language models acted as autonomous research agents under human direction across every category the venue's policy names: generating and selecting hypotheses, writing the pre-registration, implementing the stimuli, the measurement and the analysis code, running the simulations, interpreting results, and adversarially reviewing drafts. Human direction set the goal, the constraints and the standard of evidence; the authors are responsible for every claim.

S1 Pre-registration addendum for E1 and E2, as filed

Reproduced exactly as filed, including passages that respond to our own adversarial review rounds — this work was reviewed internally by independent agents before submission, and a numbered "round" in the text below refers to one of those, not to any venue's process. The document is verified against its committed hash at build time, so it is reproduced rather than edited.

Pre-registration addendum — E1 (quote vs indirection) and E2 (prior-unavailable control)

Written **before any response to the new rule forms exists**. Timestamped by the commit that adds this file. E1 and E2 are confirmatory; E3 is exploratory and makes no confirmatory claim.

0. What this resolves

The submitted paper concedes, in its own abstract, that it cannot separate two accounts of why ‘J_X‘ changes sign between R0 and R1:

> "…the latter being what a fallback to the world-knowledge answer would also produce, **which we > cannot separate here and do not claim to**."

R0 → R1 changes two things at once: the compliant answer's **string** leaves the rule, and an **indirection step** (label → answer) is added. This addendum adds the missing cell so the two factors are crossed rather than confounded.

	string present in rule	indirection
R0	yes	no
R1	no	yes
R1Q	**yes**	**yes**

1. Rule forms, verbatim

The outer frame is unchanged from ‘stimuli_e1.py‘ and identical across all forms:

‘‘‘
Question: {q}

Options:
{opt_block}

Rule:
The answer must be one of the two listed options.
{body}

Candidate answer:
{cand}

{tail}
‘‘‘

Only ‘{body}‘ differs. Existing forms (unchanged):

‘‘‘
R0 The answer must be ”{designated_answer}”.
R1 Only the answer associated with label {dl} is permitted.
R2 The answer associated with label {ol} is forbidden.
‘‘‘

New forms:

‘‘‘
R1Q Only the answer associated with label {dl} (”{designated_answer}”) is permitted.

R1QQ Only the answer associated with label {dl} (”{designated_answer}”) is permitted; the answer associated with label {ol} (”{other_answer}”) is not.

R1-long Only the answer associated with label {dl} is permitted. An answer outside the two options listed above cannot satisfy it.
‘‘‘

‘R1Q‘ is ‘R1‘ with exactly one insertion: ‘ (”{designated_answer}”)‘ after the label.

R1-long is a length control, and is semantically inert by construction: its added sentence restates the domain sentence already present in every form and supplies no information about which label is designated or how to resolve a label. Measured body lengths over the 160 items: R1 53.0 chars (sd 0.0), R1Q 63.9 (1.7), R1QQ 117.9 (2.7), R1-long 119.0 (0.0) — R1-long sits **+1.1 chars** from R1QQ’s mean, so a quote effect that is really a length effect must appear in R1-long too.

2. Estimands

Computed on the same units, same clustering, same estimator as the submitted paper.

- **Pure quote effect** ‘Q_quote = L_X(R1) − L_X(R1Q)‘. Indirection identical, differs only by the quoted string.
- **Pure indirection effect** ‘Q_ind = L_X(R1Q) − L_X(R0)‘. String present in both, differs only by the indirection step.
- **Length control** ‘Q_len = L_X(R1) − L_X(R1-long)‘. Expected 0 under every account we are testing; a non-zero value invalidates the quote reading and is reported as such.
- **Binding vs presence** ‘Q_bind = L_X(R1Q) − L_X(R1QQ)‘, with ‘p01‘ reported separately.

Each is reported with its aligned-cell counterpart, and **‘e10‘ and ‘e01‘ are reported separately** rather than only their difference.

Why option order is not crossed. ‘rule_form‘ is crossed with ‘base_id‘: the same item appears under every rule form with the same letters, the same option order and the same candidate strings, so every estimand above is a **within-item contrast** and option order is held fixed on both sides of each difference. Order is imbalanced in the bank (true_first 13,392 / false_first 9,648) but cannot enter a within-item difference. We therefore do **not** double the call budget, and instead report all four estimands **stratified by option order** as a moderator analysis, using the existing 9,648 false-first rows.

3. Sampling and inference

- Clustering unit: **construction group** (40 groups), unchanged from the submitted paper.
- Intervals: stratified paired cluster percentile bootstrap over groups, same resamples and seed as the primary analysis.
- Wording: **main wording only** (W0) for the new forms.
- Decoding, parser and oracle: identical to ‘exp_e1_main_audit‘; any deviation goes in the deviations table.

4. Decision thresholds, fixed now

Format follows the existing C3 condition.

- **E1 primary.** ‘Q_quote‘ is **established** iff the model-average interval lower bound ‘> 0.05‘ **and** at least **3 of 4** models individually have an interval lower bound ‘> 0‘.

- **E1 length control.** If ‘Q_len‘ excludes zero in the same direction as ‘Q_quote‘, the quote reading is **withdrawn** and the effect is reported as a rule-length effect.
- **E2 primary.** On the relation-only branch, quote effect ‘= BA(R1Q) − BA(R1)‘ is **established** iff its interval lower bound ‘> 0.05‘ and at least 3 of 4 models individually exclude zero. Balanced accuracy is reported, and accuracy is additionally split by whether the candidate is the permitted one.
- **Below 10 construction groups in any subset, no branch label is read** — same rule the relation-verified subset already uses.

5. Result branches, and the paper each one produces

Written now so that no narrative is chosen after seeing the numbers.

branch	result	paper
A	‘Q_quote‘ established in E1 **and** E2	Quoted rules inflate compliance scores through a mismatch-rejection shortcut, and it is not prior fallback. Title and abstract as planned.
B	‘Q_quote‘ established in E1, not in E2	The quote shortcut operates only where a prior conflict exists — an interaction. Title changes to name the interaction.
C	‘Q_quote‘ not established, ‘Q_ind‘ dominates	Main line becomes **label indirection triggers prior fallback**, so label rewrite controls are not clean substitutes for quoted rules. Title and abstract rewritten. This is a publishable finding and is not a f
D	Models disagree in direction	Scope the claim to the models where it holds, report the between-model difference as the finding, and read it against E3’s scale axis.

‘Q_len‘ non-zero overrides A and B and sends the paper to a length-effect framing.

6. E3

Exploratory. Adds larger models on main wording and R0/R1/R1Q only, plus the relation-only branch
as a rule-reading gate. No threshold governs it and no confirmatory claim rests on it. It answers ”does the effect exist above 8B”, not ”how does it scale”, and the paper will say so.

---

7. Amendment, recorded 2026-09-16 22:55, before any E2 response exists

E2 wording. Section 3 fixes ”main wording only (W0)” for the new forms. That was written for E1 and cannot apply to E2: the existing relation-only arm has no W0 at all — it is built on the paired WN/WC blocks, because the relation-only prompt removes the question and W0 is the earlier study’s question wording. E2 therefore runs **WN and WC, both**, matching the existing relation-only arm exactly so its R1 numbers are comparable to the ones already reported.

This costs calls rather than buying a result: E2 is 7,680 stimuli x 4 models = 30,720 calls. We record the choice here rather than pick a wording after seeing which one is kinder.

E2 pseudoword variant. Both variants (‘real‘, ‘pseudo‘) are collected. The pseudoword arm replaces both candidates with invented strings drawn deterministically from a fixed pool, so a quote effect cannot ride on any association the real answer string already carries. Pre-registered reading: if the quote effect holds on ‘real‘ but vanishes on ‘pseudo‘, the effect is about familiar strings rather than about string presence, and the paper says that instead.

---

8. Addendum 2 — R1Qneg, filed before any response to this form exists

Why. R1QQ prints both strings, but the non-designated one appears inside a negation (”...is not”), so a model could be parsing the negation rather than the binding. A rule that quotes **only the forbidden answer** separates the two accounts without that confound.

Form. R1 with one insertion, mirroring R1Q but on the other label:

‘‘‘
R1Qneg Only the answer associated with label {dl} is permitted; the answer associated with label {ol} (”{other_answer}”) is not.

‘‘‘

The permitted answer’s string never appears. The forbidden answer’s string does.

Estimands and pre-declared readings.

- ‘P_neg01 = p01(R1) - p01(R1Qneg)‘, and ‘P_neg10 = p10(R1) - p10(R1Qneg)‘.
- **Mismatch-rejection against any printed string:** the model rejects a candidate that does not match the printed string. Here the printed string IS the non-designated candidate, so it should be *endorsed* more often: ‘p01‘ RISES, i.e. ‘P_neg01 < 0‘, and the interval excludes zero.
- **Binding to the permitted label:** the permitted string is absent, so there is nothing to match against and behaviour should track R1: ‘P_neg01 ~ 0‘, interval containing zero.
- **Retrieval-cost account** (the objection round 2 raised): R1Qneg supplies no help in retrieving the *permitted* label’s string, so it should behave like R1 on ‘L_X‘ — ‘L_X(R1) - L_X(R1Qneg) ~ 0‘. If instead it moves as much as ‘Q_quote‘ does, the effect is about printed strings generally rather than about the permitted binding.

Threshold. Same form as before: an effect is established iff the model-average interval lower bound exceeds 0.05 in the predicted direction and at least 3 of 4 models individually exclude zero. Below 10 construction groups in any subset, no branch label is read.

Both outcomes are reportable. If ‘p01‘ rises, the mechanism is closer to plain string matching than the binding reading in Section 4.2 allows, and that section is rewritten. If it does not move, the binding reading survives a test designed to break it.

9. Addendum 3 — R1Qx, the inconsistent quote, filed before any response to this form exists

Why. Every form run so far makes the rule’s label binding and the printed string agree, or prints the string inside a negation. In all of them the rule’s designation *is* the ground truth, so a printed string that the model merely copied could never produce an error and a string effect could never register as inflation. R1Qneg was meant to separate the two accounts and did refute one of them, but its printed string is explicitly marked forbidden, so rule and cue still point the same way. This form makes them point in **opposite** directions.

Form. R1Q with one substitution: the quoted string in the permitted slot is the string of the **non-designated** candidate.

‘‘‘
R1Qx Only the answer associated with label {dl} (”{other_answer}”) is permitted.
‘‘‘

The rule is internally inconsistent by construction: the label binding designates one answer, the quoted string names the other. Everything else -- question, both candidates, option order, letters, wording -- is identical to R1Q, so this is a within-item contrast like all the others.

Scoring. The designated answer is the one the **label binding** names, exactly as in R1Q and R1. That choice is made here, in advance, and is the only choice that keeps the contrast with R1Q within-item: if we scored against the quoted string instead, the two forms would not share a ground truth and no difference between them would be interpretable.

Estimands and pre-declared readings.

- ‘X_desig = (p11 + p10)/2 - (p01 + p00)/2‘ under R1Qx: endorsement of the label-designated candidate minus endorsement of the other one. Under R1Q this quantity is strongly positive.
- ‘Q_incons = L_X(R1Qx) - L_X(R1Q)‘: what an inconsistent quote costs, scored against the label.
- **String cue governs.** The model endorses whatever string the rule prints, so it endorses the non-designated candidate: ‘X_desig < 0‘, interval excluding zero. This is the reading the submitted title asserts, and it has never been tested against a case where the two disagree.
- **Label binding governs.** The model resolves the label and ignores the inconsistent string: ‘X_desig > 0‘, of the same order as under R1Q, and ‘Q_incons ~ 0‘.
- **Neither governs.** ‘X_desig ~ 0‘, or a rise in unreadable replies: the inconsistency is detected and the model refuses or hedges. Unreadable rates are reported per form either way.

Threshold. As before: an effect is established iff the model-average interval lower bound exceeds 0.05 in the predicted direction and at least 3 of 4 models individually exclude zero.

Responsiveness is reported beside every per-model count, using the same 0.10 conflict-cell floor declared in Section 8's analysis, and both the screened and unscreened counts are printed.

Both outcomes are reportable, and they name different papers.

- 'X_desig < 0': a printed string can override an explicit label binding. The string-matching reading is restored, the title stands, and the effect is an artefact in the strong sense.
- 'X_desig > 0': models follow the rule when the string disagrees with it, so the printed string is not a cue that overrides the rule. The effect measured by 'Q_quote' is then the cost of **resolving** a rule whose extension is not written out, the title must stop saying "string match", and the paper's claim becomes that a compliance score is partly a measure of how explicitly the rule is written. The abstract already carries the number; only its interpretation changes.

We record here, before running, that the second outcome is the one the R1Qneg result leads us to expect, and that it requires us to retitle the paper.

10. Addendum 4 — the benchmark rule-form audit, coding rule fixed before counting

Why. Two consecutive review rounds named the same gap: the paper asserts that evaluations write rules that print the answer, and never evidences it. This addendum fixes the coding rule before any count is produced, so that the count is a measurement rather than an impression.

Evidence collection. Rule and instruction templates were collected from primary sources -- the benchmark's own repository, dataset card, or paper -- by a process that recorded verbatim text and URLs and performed **no classification**. Text that could not be quoted is recorded as 'null' or explicitly prefixed 'PARAPHRASE:'; neither is counted. The collected file is 'experiments/data/benchmark_rule_forms.json' and ships with the artifact.

Unit of analysis. For IFEval the unit is the *verifiable-instruction type*: the 24 description templates in the published 'instructions.py'. Every other benchmark contributes its rule template as one unit. The IFEval count is the primary because it is the only population that is both enumerable and published in full.

Coding rule, fixed here. A rule template is coded

- **PRINTS** iff it contains a placeholder whose instantiated value is a literal string that must appear verbatim in a compliant response (for example '{keyword}', '{ender}', '{starter}');
- **PRINTS-NEGATED** iff it prints such a string whose *absence* is what compliance requires;
- **PROPERTY** iff compliance is decided by a property of the response -- a count, a format, a language, a casing -- with no string in the rule for the response to match;
- **UNDETERMINED** iff no rule text is published.

The coding of each unit is committed as a table with the template text beside it, so a reader can check every assignment against the quoted source rather than taking the count on trust.

What the count can and cannot support. A high PRINTS share establishes that the rule form this paper manipulates is the form published benchmarks use. It does **not** establish that those benchmarks' scores are inflated: their items are not ours, we ran no model on them, and the effect size measured here transfers to them only under an assumption we do not test. The paper will say exactly that.

11. Addendum 5 — R1pos, separating "a string to match" from "a label to dereference"

Why. Every low-error form so far (R0, R1Q, R1QQ, R1Qneg) prints a candidate string, and every high-error form (R1, R1-long) requires resolving an arbitrary letter to an option. Two accounts therefore fit all six forms:

- **String matching.** The decision is a comparison against whatever candidate strings the rule prints, in the polarity of the slot they appear in. Absent a string, the model falls back on the label binding and does it poorly.
- **Dereference cost.** The models cannot reliably resolve an arbitrary label to an option, and every form whose extension is computable *without* that resolution scores better. Printing a string is one way to make it computable; it is not the operative variable.

R1Qx does not separate them: a model that cannot dereference would also follow the string when the two disagree. This form does, by making the extension computable **without printing any candidate string and without an arbitrary label**.

Form. R1 with the designation made positional:

‘‘‘
R1pos Only the answer listed {first|second} above is permitted.
‘‘‘

‘{first|second}‘ names the position of the designated option in the options block, which is already fixed per item by the frozen stimulus construction. No candidate string appears anywhere in the rule, and no arbitrary letter has to be resolved.

Estimands and pre-declared readings.

- ‘Q_pos = L_X(\text{R1}) - L_X(\text{R1pos})‘: what removing the arbitrary label buys while keeping every candidate string out of the rule.
- **String matching predicts ‘Q_pos ~ 0‘**: nothing is printed for the model to compare against, so behaviour should track R1, and ‘Q_pos‘ should be small beside ‘Q_quote = 0.153‘ measured on the same items.
- **Dereference cost predicts ‘Q_pos‘ of the same order as ‘Q_quote‘**: the resolution step is what was expensive, and R1pos removes it.
- Intermediate values are reportable as intermediate. If ‘Q_pos‘ lands between the two, both factors operate and the paper says so, with ‘Q_pos / Q_quote‘ as the reported split.

Threshold. An account is established iff the model-average interval lower bound exceeds 0.05 in its predicted direction and at least 3 of 4 models individually exclude zero, and the screened count over responsive checkpoints is reported beside the unscreened one.

Also filed here: the wording extension. The crossed forms were run on the bridge wording W0 only, while the R0/R1 comparison has all three. We will run R1Q on the two paired wordings WN and WC and report ‘Q_quote‘ per wording. This is a robustness check with no threshold attached: the registered claim remains the W0 estimate, and a disagreement across wordings would be reported as a limit on the estimate rather than as a new result.

Both outcomes are reportable, and they name different papers. If ‘Q_pos ~ 0‘, the printed string is the operative variable and the paper's title stands. If ‘Q_pos‘ matches ‘Q_quote‘, the operative variable is the dereference and the paper becomes ”rules requiring an extra dereference deflate small-model compliance scores” --- branch C of Section 5's table, a different title and a different abstract, which we would write.

12. Addendum 6 — R1Qx with no prior available, filed before the run

Why. In R1Qx on the main bank, the string the rule quotes is the non-designated candidate, which is the baseline-\emph{correct} answer exactly in the conflict condition and the incorrect one in the aligned condition. Endorsement could therefore track the prior rather than the string. The four cells of the run already separate these --- a prior account cannot explain endorsement of the true, label-designated candidate collapsing when the rule quotes the other answer --- but the clean test is to remove the prior entirely.

Form. R1Qx on the relation-only branch: no world-knowledge question, two labelled options, the rule binding one label and quoting the other option's string. Includes the pseudoword variant, where neither option string has any prior association at all.

Estimand and pre-declared readings. ‘A_label‘ is accuracy against the label binding.

- **String matching.** ‘A_label‘ falls far below chance, because the model endorses the quoted (non-designated) option: the error is systematic, not noise.
- **Prior following.** With no question and no true answer there is nothing for a prior to prefer, so on the real-word variant ‘A_label‘ should be at chance and on the pseudoword variant certainly so.
- **Label binding.** ‘A_label‘ stays high, as it is under R1Q on this branch.

Threshold. The string-matching reading is established iff the model-average ‘A_label‘ interval lies entirely below \literal{0.5} and at least 3 of 4 models individually do, on the pseudoword variant as well as the real one.

13. Addendum 7 — R1Qdesc, separating ”a matchable string” from ”the extension written out”

Why. Across every form run so far, two things move together: whether the rule prints a candidate string, and whether the rule writes out its own extension. The four string-printing forms occupy the four best ‘L_X‘ ranks and the three stringless forms the three worst, and no contrast we have separates the factors. R1pos was filed to break the confound and does not: it substitutes an ordinal reference for an alphabetic one rather than making the extension computable from the rule text, and its ‘p01‘ shows the models did not resolve it.

Form. R1Q with the quotation replaced by a definite description of the same answer, taken from the frozen bank: every candidate in this design is the true answer to some base question in its construction group, so the description is that question, and no candidate string appears in the rule.

‘‘‘
R1Qdesc Only the answer associated with label {dl} (the answer to: {question naming that
        answer}) is permitted.
‘‘‘

Estimands and pre-declared readings.

- ‘Q_desc = L_X(R1) - L_X(R1Qdesc)‘: what writing the extension out buys when nothing in the rule is matchable against a candidate.
- **Matchable string.** The operative variable is a string the candidate can be compared against, so a description buys nothing: ‘Q_desc ~ 0‘, and ‘L_X(R1Qdesc)‘ sits at ‘L_X(R1)‘.
- **Extension written out.** Any rule that determines its answer within its own text is easier, matchable or not: ‘Q_desc ~ Q_quote‘, and ‘L_X(R1Qdesc)‘ sits at ‘L_X(R1Q)‘.
- Intermediate values are reportable as intermediate, with ‘Q_desc / Q_quote‘ as the split.

Threshold. An account is established iff the model-average interval lower bound exceeds 0.05 in its predicted direction and at least 3 of 4 models individually exclude zero; the screened count over responsive checkpoints is reported beside the unscreened one.

Both outcomes are reportable, and they name different papers. If ‘Q_desc ~ 0‘, the title’s ”string match” is established against its closest rival and the paper stands. If ‘Q_desc ~ Q_quote‘, the operative variable is explicitness rather than matchability, the title must say so, and the paper becomes a claim about how much of a rule’s extension the evaluation writes out. We record before running that the second outcome requires the retitle.

A separate correction filed at the same time. Section 5’s branch table assigned each outcome a consequence in advance, and the outcome we got --- ‘Q_quote‘ established in E1, E2 below its margin --- is branch B, whose stated consequence is that the title names the interaction. We did not do that, and the deviations table will say so and give the reason: the interaction is a quantity we decided not to estimate, because the two arms differ in scale, sample and wording.

Amendment to Addendum 7, filed before the run. Two properties of the implementation, written down now because they bear on how each outcome may be read.

1. The description is only non-degenerate in the **conflict** condition. In the aligned condition the designated answer is the true answer to the question already printed above the rule, so the description restates that question. ‘L_X‘ is computed entirely from the conflict cells, which is the registered estimand; the aligned cells of this form are degenerate by construction and will not be interpreted.
2. The two readings are not symmetric in what they license. ‘Q_desc ~ Q_quote‘ cleanly kills the matchable-string reading. ‘Q_desc ~ 0‘ is weaker evidence for it, because resolving the description requires a second world-knowledge step, so a description could fail to help for a reason unrelated to matchability. We will report that asymmetry with the result rather than claiming the stronger reading from the weaker branch.

14. Addendum 8 — R1Qsep, the quotation detached from the label

Why. In R1Qx the contradicting string sits inside a parenthesis attached to the label it contradicts: ‘label M ("Paris")‘. A model that reads the parenthesis as *fixing what the label denotes* is applying a coherent rule, not overriding one, and that account reproduces every number we reported. It is not among the three the previous addendum enumerated, and it is the reading under which our title's third clause is wrong.

Form. The same inconsistency, with the string moved out of the label's apposition into its own sentence, so it cannot be read as a gloss on the label:

‘‘‘
R1Qsep Only the answer associated with label {dl} is permitted.
        The permitted answer is "{other_answer}".
‘‘‘

As in R1Qx, the quoted string is the non-designated candidate and scoring is against the label binding, fixed here before the run.

Estimands and pre-declared readings.

- ‘X_desig(R1Qsep)‘, as in R1Qx: endorsement of the label-designated candidate minus the other.
- **Apposition reading.** The string governed only because it glossed the label. Detached, it is a second sentence in conflict with the first, so behaviour should move back toward the label: ‘X_desig‘ rises substantially above its R1Qx value and its interval excludes that value.
- **The string governs.** Position in the sentence is not what mattered: ‘X_desig(R1Qsep)‘ is close to ‘X_desig(R1Qx)‘ and remains negative.

Threshold. The string-governs reading survives iff ‘X_desig(R1Qsep)‘ is negative with its model-average interval excluding zero and at least 3 of 4 models individually negative. The apposition reading is established iff ‘X_desig(R1Qsep) - X_desig(R1Qx)‘ exceeds 0.05 with its interval excluding zero.

Both outcomes are reportable. If the apposition reading holds, the title's third clause is withdrawn and R1Qx is reported as showing that a parenthetical gloss overrides the label it glosses --- a narrower and still reportable finding. If it does not, the override claim stands against the sharpest alternative raised against it.

15. Addendum 9 — R1Qdescx, the contradicting description

Why. The disagreement control has been run three times against the quotation (R1Qneg, R1Qx, R1Qsep) and never once against the **description**, which is the factor the title now names. As it stands, R1Qdesc does not remove the matchable string: it puts it one retrieval step away, and on this bank the roster resolves those retrievals at ceiling. "Dereference the description, then match" predicts exactly what was observed --- ‘Q_desc‘ slightly below ‘Q_quote‘, the gap being retrieval cost --- which the paper currently labels matchability. This form separates them.

Form. R1Qdesc with the description pointing at the **other** candidate:

‘‘‘
R1Qdescx Only the answer associated with label {dl} (the answer to: {question whose true
          answer is the NON-designated candidate}) is permitted.
‘‘‘

The label designates one candidate and the description identifies the other, exactly as R1Qx does with a quotation. Scoring is against the label binding, fixed here before the run, so the contrast with R1Qx is like for like. As in R1Qdesc the description is degenerate in the aligned condition and ‘L_X‘ is computed from the conflict cells, which is the registered estimand.

Estimands and pre-declared readings.

- ‘X_desig(R1Qdescx)‘, as in R1Qx: endorsement of the label-designated candidate minus the other.
- **The description governs, like the quotation.** Then explicitness and matchability are one mechanism a retrieval step apart: ‘X_desig < 0‘, interval excluding zero, on at least 3 of 4.
- **The label governs.** A description does not override a binding the way a printed string does: ‘X_desig > 0‘, of the order it takes under R1Qdesc.

Threshold. Established iff the model-average interval lower (or upper) bound clears zero by more than 0.05 in the predicted direction and at least 3 of 4 models individually agree, with the screened count over responsive checkpoints reported beside the unscreened one.

Both outcomes are reportable, and they name different papers. If the description governs, the paper's first clause is not "explicitness" but "any resolvable pointer to an answer, printed or described, overrides the rule's own designation" --- one mechanism, and the title must say so. If the label governs, explicitness and matchability are genuinely different factors and the decomposition stands as written. We record before running that the first outcome merges the two components the paper currently separates.

16. Addendum 10 — the property rule, which no form here has ever been

Why. The audit filed as addendum 4 codes every published verifiable instruction into three kinds: 'PRINTS' (the rule contains a literal string the response must carry), 'PRINTS-NEGATED', and **'PROPERTY'** (compliance is decided by a property of the response --- a count, a format, a language --- with no string in the rule for the response to match). 'PROPERTY' is the largest of the three, 13 of the 24 IFEval types. **No 'PROPERTY' rule has ever been run here.** All twelve forms designate a member of the options block: by quoting it (R0, R1Q, R1QQ, R1Qx, R1Qsep), by an arbitrary letter (R1, R1-long, R1Qneg), by an ordinal (R1pos), or by a description that still hangs off a letter (R1Qdesc, R1Qdescx).

The consequence is that 'Q_quote' is 'PRINTS' minus **label indirection into an options block**, and label indirection into an options block is a construct this paper built to serve as its own control; the audit reports finding no such rule in the field. The field's own contrast is 'PRINTS' against 'PROPERTY', and this paper's estimate of it is currently empty. That gap is the distance between the paper's §5, which says "most of what a conflict-only score measures **here** is how the rule was written", and its title, which drops the *here*.

Forms. Two, because one property is one capability and a single-property result could not distinguish the rule form from the arithmetic it happens to demand. Both leave the options block byte-identical to every other form, so the contrast stays within item.

```
RPROPN   Only an answer of exactly {n} letters is permitted.
         n = the count of alphabetic characters in the designated answer. Requires the two
         candidates to differ in that count; items where they do not are dropped, and the
         dropped count is reported per group.

RPROPV   Only an answer {phi} is permitted.
         phi is the first predicate, in this fixed order, that holds of the designated candidate
         and fails of the other: "that is a single word"; "that is more than one word long";
         "that begins with a vowel"; "that ends in a vowel"; "of exactly {n} letters".
         The order is fixed here, before any response exists, so the per-item choice is
         deterministic and not a search.
```

Neither rule contains either candidate's string, and neither names a label or a position. Scoring is unchanged: 'C' is whether the candidate satisfies the rule, 'T' is whether it is the true answer, and the designated candidate is the same one every other form designates, so 'L_X' is computed from the same conflict cells as everywhere else.

Estimands.

- 'Q_prop = L_X(RPROPN) - L_X(R0)' and the same against 'RPROPV': what a property costs relative to a rule that prints the answer.
- 'Q_propind = L_X(R1) - L_X(RPROPN)': what a property costs relative to arbitrary-label indirection --- the contrast the field actually faces.

The two live accounts predict opposite signs, which is why this is worth the run.

- **Matchability** (this paper's account): what makes a rule cheap is that the compliant string is printed where it can be compared against the candidate. A property is not a string, so it should behave like R1, not like R0: 'Q_prop > 0', of the order of 'Q_quote'.
- **Cross-reference cost** (the account raised against this paper): what makes a rule expensive is

having to go back to the options block and resolve a symbol. A property rule requires no such return --- "does this candidate have exactly five letters" is answerable from the candidate alone --- so it should behave like R0: ‘Q_prop ~ 0‘, and ‘Q_propind ~ Q_quote‘.

Thresholds, fixed before the run. ‘Q_prop‘ is established as positive iff the model-average interval lower bound exceeds +0.05 and at least 3 of 4 checkpoints are individually positive with intervals excluding zero, the screened count over responsive checkpoints reported beside the unscreened one. It is established as null iff the model-average interval is contained in (-0.05, +0.05) on **both** property forms.

Both outcomes are reportable, and they name different papers.

- If ‘Q_prop > 0‘: the paper's factor survives contact with the field's own alternative. The abstract's claim that the cost attaches to a rule form a large minority of published instructions use is then measured rather than asserted, and ‘Q_propind‘ gives the field contrast a number for the first time.
- If ‘Q_prop ~ 0‘: **the paper's stated factor is wrong and the title must change.** What is cheap is not that the answer is printed but that the rule is self-contained, and the expensive thing is the options-block dereference, which is a property of this harness and of no published suite. The empirical half then supports a narrower claim --- conflict-only scores on *multiple-choice-style* compliance harnesses are dominated by how the rule points at its referent --- and every sentence generalising past this harness comes out. We record before running that this outcome removes the paper's present title.

16.1 Pre-run amendment to addendum 10 (2026-09-17, before any response existed)

The four predicates registered above were checked against the bank before running, and they separated the two candidates on only 27% of designation instances: the remaining 73% fell through to the count clause, which would have made ‘RPROPV‘ a noisier copy of ‘RPROPN‘ rather than a second property. Two forms that are one form cannot tell a rule form from the arithmetic it demands, which is the entire reason addendum 10 asked for two. The registered list is therefore replaced, now, before any model has seen a property rule:

‘‘‘
1. that is a single word
2. that is more than one word long
3. that begins with a vowel
4. that does not begin with a vowel
5. that ends in a vowel
6. that does not end in a vowel
7. in which some letter appears more than once
8. in which no letter appears more than once
9. whose first letter falls between A and M in the alphabet
10. whose first letter falls between N and Z in the alphabet
    fallback: of exactly {n} letters
‘‘‘

Same rule as before: the first predicate in this order that holds of the designated candidate and fails of the other. On the 320 designation instances of the audit split this gives **266 (83.1%) a non-count predicate, 38 (11.9%) the count fallback, and 16 (5.0%) no separating predicate at all**; those 16 are dropped, per group, and the dropped count is reported with the result. These three numbers are recorded here before the run so that the composition of ‘RPROPV‘ cannot be adjusted after seeing how it came out.

This amendment adds negated predicates (4, 6, 8, 10). A negated property is still a property in the audit's sense --- IFEval's ‘{no_comma}‘ is one --- and it prints no candidate string, names no label and fixes no position, which are the three things that make a form ‘PROPERTY‘ here.

17. Addendum 11 — can these checkpoints evaluate the property at all?

Why, and this one is a correction. Addendum 10's result was written up with the claim that the property-rule collapse is not a difficulty effect, and the evidence offered was aligned-cell error: ‘L_A‘ is 0.052 and 0.058 under the property forms, so — the argument went — the models can apply the property. **That argument is invalid, and this paper says so elsewhere.** In an aligned cell the compliant candidate *is* the true answer, so a model that ignores the rule entirely and answers

from the facts scores near-zero aligned error. ‘L_A‘ therefore cannot separate ”applies the property” from ”answers from truth”, which is exactly the objection this paper raises against aligned-cell controls in its own appendix. The sentence was wrong when written and is withdrawn here rather than defended.

What the argument needs is a condition with **no truth to fall back on**. The relation-only branch is that condition: no question, no true answer, both options either real words or invented ones, and the rule is the only thing that decides. Accuracy there is a direct measure of whether these checkpoints can evaluate the property.

Form. ‘RPROPN‘ and ‘RPROPV‘, exactly as addendum 10 and 16.1 define them, built on the relation-only branch alongside the R0/R1/R1Q forms already run there, on both the real-word and pseudoword variants. Items where no registered predicate separates the two options are dropped as before, and the dropped count is reported per group.

Estimand. ‘A_rule(PROPERTY)‘, balanced accuracy against the rule, per variant and per checkpoint — the same quantity the relation-only arm already reports for the label forms, so the comparison is like for like.

Pre-declared readings, and they are exclusive.

- **The checkpoints can evaluate the property.** ‘A_rule‘ is at or above the level the label forms reach on the same branch, on at least 3 of 4 checkpoints. Then the main-bank collapse is not competence: the models can decide these properties when the rule is the only thing to go on, and fail to apply them when a fact competes. ‘Q_prop‘ stands as a statement about rule form.
- **They cannot.** ‘A_rule‘ is at or near chance on the property forms while the label forms on the same stimuli are not. Then ‘Q_prop‘ is not separable from property-evaluation difficulty, and **the claim that a property rule is the most expensive rule form is withdrawn**: what would remain is that these checkpoints cannot be scored on property rules at all, which is a statement about the instrument and the checkpoints, not about what a compliance score measures.

Threshold. The first reading is established iff the model-average ‘A_rule‘ on the property forms has its interval lower bound above 0.6 on the pseudoword variant — the variant with no lexical association to lean on — and at least 3 of 4 checkpoints individually exceed 0.6. The second is established iff the model-average interval contains 0.5 on the pseudoword variant. Anything else is reported as indeterminate and the ‘Q_prop‘ claim is narrowed to a statement conditional on property competence, with the numbers given.

Recorded before running: the second outcome removes the property result from the abstract and returns the paper to the scope round 21 criticised it for. We run it because the first outcome is the only thing that licenses the sentence we withdrew, and an unlicensed sentence is worse than a narrower paper.

17.1 Also registered here: does the cost track predicate difficulty?

From the addendum-10 logs already collected, ‘L_X‘ is broken out by which registered predicate the rule used (single word, vowel position, repeated letter, alphabet half, letter count). If the cost is a rule-form effect it is flat across predicates; if it is arithmetic competence it concentrates on the count clause, which is the only predicate requiring counting. This is a descriptive breakdown, not a threshold, and it is reported whichever way it falls.

S2 Per-cell results for the design reported in the main text

Every cell of every rule form, for the wording the new forms were run on (W0), emitted from the response logs rather than typed — including R1Qneg, R1Qx and R1pos, the three forms that carry the mechanism claim. p_{ct} is the complete-case endorsement rate in cell $(C{=}c, T{=}t)$; L_A and L_X are the aligned and conflict operational errors defined in Section 3.1. Every cell holds 160 responses and no response was unreadable, so the complete-case rate and the unconditional rate coincide throughout and no imputation arises.

model	rule	p_{11}	p_{10}	p_{01}	p_{00}	L_A	L_X	n/cell
Gemma-3-4B	R0	1	0.212	0	0	0	0.394	160
	R1	0.994	0.100	0.269	0	0.003	0.584	160
	R1-long	0.988	0.100	0.419	0	0.006	0.659	160
	R1pos	0.981	0.106	0.394	0	0.009	0.644	160
	R1Q	0.994	0.237	0.006	0	0.003	0.384	160
	R1Qdesc	1	0.438	0.200	0	0	0.381	160
	R1Qneg	1	0.169	0.050	0	0	0.441	160
	R1QQ	1	0.412	0.013	0	0	0.300	160
	R1Qsep	0.050	0	0.856	0.037	0.494	0.928	160
	R1Qx	0.075	0	0.994	0.231	0.578	0.997	160
Llama-3.2-3B	R0	0.988	0.025	0.006	0	0.006	0.491	160
	R1	1	0.081	0.019	0	0	0.469	160
	R1-long	1	0.181	0.219	0	0	0.519	160
	R1pos	0.981	0.006	0.581	0	0.009	0.787	160
	R1Q	1	0.056	0	0	0	0.472	160
	R1Qdesc	1	0.175	0.050	0	0	0.438	160
	R1Qneg	1	0.594	0	0	0	0.203	160
	R1QQ	1	0.481	0	0	0	0.259	160
	R1Qsep	0.119	0.037	0.681	0	0.441	0.822	160
	R1Qx	0.181	0.025	0.781	0.006	0.412	0.878	160
Qwen3-4B	R0	1	0.125	0	0	0	0.438	160
	R1	0.988	0.062	0.331	0	0.006	0.634	160
	R1-long	0.988	0.081	0.388	0	0.006	0.653	160
	R1pos	0.994	0.013	0.656	0	0.003	0.822	160
	R1Q	1	0.494	0	0	0	0.253	160
	R1Qdesc	0.988	0.131	0.006	0	0.006	0.438	160
	R1Qneg	0.988	0.637	0	0	0.006	0.181	160
	R1QQ	1	0.775	0	0	0	0.113	160
	R1Qsep	0	0	0.988	0.025	0.512	0.994	160
	R1Qx	0.075	0.006	0.994	0.069	0.497	0.994	160
Qwen3-8B	R0	0.994	0	0	0	0.003	0.500	160
	R1	0.988	0.006	0.062	0	0.006	0.528	160
	R1-long	0.975	0.013	0.087	0	0.013	0.537	160
	R1pos	0.981	0.006	0.306	0	0.009	0.650	160
	R1Q	0.994	0.013	0	0	0.003	0.494	160
	R1Qdesc	0.988	0	0	0	0.006	0.500	160
	R1Qneg	0.988	0.013	0	0	0.006	0.494	160
	R1QQ	1	0.044	0	0	0	0.478	160
	R1Qsep	0	0	0.469	0	0.500	0.734	160
	R1Qx	0	0	0.900	0	0.500	0.950	160

The contrasts in Section 4.2 are differences of the L_X column taken within model and within item. Reading down that column for one model gives each of them directly: Q_{quote} is R1 minus R1Q, Q_{ind} is R1Q minus R0, Q_{len} is R1 minus R1-long, and Q_{bind} is R1Q minus R1QQ. The p_{01} column carries the ordering discussed there: it sits at or below 0.013 in every row whose rule prints the permitted answer (R0, R1Q, R1QQ) and rises to between 0.019 and 0.419 in every row whose rule does not (R1, R1-long), across all four checkpoints.

Interval coverage, and which intervals it covers. The pre-registration's simulation was run on the C1 estimator at a nominal 0.95 and puts its group-clustered percentile coverage just under that. The crossed-form contrasts — Q_{quote}, Q_{ind}, Q_{desc}, X_{desig} and the rest — are now reported at the same nominal 0.95 (D14), but their coverage was not simulated: the study was built for the C1 estimator. Sharing a nominal level is not sharing a coverage study, and we state that rather than transfer the C1 figure to intervals it was not computed for. What the simulation supports is that the C1 threshold is, if anything, slightly optimistic; for the crossed forms the honest statement is that the estimator is the same construction on the same clusters, and that no coverage study of it exists here.

S3 Power and coverage simulation at the frozen design

Coverage, bias and the declared rule's firing rate were simulated at the frozen design before the confirmation run, over 2000 datasets per scenario (Appendix S3). Coverage of the C1 estimator's intervals runs 0.928–0.948 against a nominal 0.95 — slightly under, so that threshold is if anything slightly optimistic, and we report it rather than correcting after seeing it. The crossed-form intervals are nominal 0.95 as of D14 and were not part of this simulation.

Coverage, bias and the firing rate of the declared rule were simulated at the frozen design before the confirmation run, under group heterogeneity, aligned cells near the floor, one strongly deviant model of four, and unreadable rates depending on condition and wording: 2000 datasets per scenario, 2000 cluster resamples each, 40 groups. Coverage of the C1 estimator's intervals runs 0.928–0.948 against a nominal 0.95; it is slightly under nominal, so that threshold is if anything slightly optimistic. The simulation was built for the C1 arm and does not cover the crossed-form contrasts; those are reported at the same nominal 0.95 since D14, but no coverage study of them exists here, and Appendix S24 says so.

S4 What would have had to be true for the aligned cells to change a selection

Both four-cell policies score a configuration by its worst cell, so the aligned cells can change a selection only when aligned error rivals conflict error. Across the twelve configurations aligned error runs 0–0.224 while conflict error runs 0.716–0.992. The aligned error at the selected configuration would have had to reach 0.721 to change the choice; it was 0.078, a margin of 0.643. This analysis is secondary and was not pre-registered. It states a condition rather than a benefit: the aligned half of the audit is worth its cost when a verifier can fail on aligned items, and on this roster it essentially cannot.

S5 The criterion-naming comparison in full

Naming the decision criterion, against a wording matched on length, structure and emphasis rather than against that study's own original question, does not selectively damage the aligned condition: $D = 0.013$ $[-0.001, 0.025]$, the registered joint rule unmet, at a power of 0.955 against the declared effect — a resolved negative, not an underpowered one (Appendix S5). The second is E2 above.

The earlier unpublished study found that naming the decision criterion selectively damages the aligned condition, comparing against that study's own original question — which differs in length, structure and emphasis as well as in naming the criterion. Against a wording matched on all of those, the effect does not reproduce: the selectivity contrast is $D = 0.013$ $[-0.001, 0.025]$, the pre-registered joint rule is not met, and the power of that rule against an effect of the declared size is 0.955, so this is a resolved negative and not an underpowered one. Coverage, bias and firing rates were simulated at the frozen design before the confirmation run; coverage of this estimator's intervals runs 0.928–0.948 against a nominal 0.95, slightly under, so this threshold is if anything slightly optimistic (Appendices S5 and S3).

That study compared a criterion-naming question against its own original question and found a large drop in the aligned contrast; that comparison differs in length, structure and emphasis as well as in naming the criterion. Against a wording matched on all of those and differing in one sentence, the effect does not reproduce: naming the criterion reduces error in both conditions, $\Delta L_A = -0.009$ $[-0.017, -0.002]$, and the selectivity contrast $D = 0.013$ $[-0.001, 0.025]$ does not exclude zero, so the pre-registered joint rule — lower bounds above 0.05 and 0.03 — is not met. Its power against an effect of the declared size is 0.955 (Appendix S3), so this is a resolved negative and not an underpowered one. The decomposition and the restricted arm are Appendix S5.

The pre-declared primary for C1 pools the permissive and prohibitive rule forms. The restricted arm using only the permissive rule form gives $D = 0.023$ $[0.010, 0.036]$; the pre-

declared primary pools both forms and we report the decomposition rather than adopting the subgroup that clears zero.

S6 Selections and test error of the four audit policies

Each policy audited the same twelve candidate configurations at equal budget and selected one; the TEST column is the balanced error of the selected configuration on the held-out slice.

policy	selected	TEST balanced error
B0 conflict scalar	Qwen3-8B|WC	0.217
B1 full H, F	Qwen3-4B|WN	0.228
B2 mixed coverage	Qwen3-4B|WN	0.228
X crossed	Qwen3-4B|WC	0.235

S7 Additivity of the two components, and the two samples it can be checked on

Both are differences of L_X, so they must sum to the R0-to-R1 cost on whatever sample they use, and they do: the new forms were run on the bridge wording alone, where $L_X(\text{R1}) - L_X(\text{R0}) = 0.098$ and $0.153 + -0.055$ reproduces it exactly. The $Q_X = 0.156$ above is a larger sample — all three wordings, averaged over the checkpoints passing Section 4.3 — so checking additivity against that figure leaves a gap of about 0.06. The gap is the wording set, not an inconsistency.

S8 Per-model conflict-cell cost of removing the quote

Per model, $Q_X = 0.291$ for Qwen3-4B, 0.252 for Gemma-3-4B and 0.111 for Qwen3-8B, but -0.028 $[-0.048, -0.011]$ for Llama-3.2-3B — resolved in the opposite direction. The pre-declared rule requires three of four, so the claim names those three and not the roster.

S9 Archive: the superseded

This section exists so that nothing is quietly dropped. Each entry names what was claimed or run, the run ids behind it, and why it was retired. None of it is deleted: the logs, the analysis code and the earlier draft's sections all ship with the artifact, and every entry below is written in the past tense because none of it is a current conclusion.

A mechanism reading, refuted in its loose form and re-established in a precise one. The submitted draft read the quoted-string effect as mismatch rejection: the verifier compares the candidate against the string the rule prints and rejects on mismatch. As stated that reading is wrong, and two pre-registered results retired it — Q_{bind} excluded zero in the direction it forbids, and R1Qneg, which quotes only the forbidden answer, lowered p_{01} where the loose reading predicts a rise. What R1Qx then established is the precise form: the comparison is against whatever string the rule prints, in the polarity of the slot it appears in. The loose reading is withdrawn; Section 4.2 states the one that survived a test built to break it. Runs:

2026-09-17T05-47-32_exp_e1q_main_audit	R1Q / R1QQ / R1LONG generation
2026-09-17T08-01-08_exp_e1qneg_main_audit	R1Qneg, the discriminating form
2026-09-17T08-21-43_analyze_e1q	contrasts, the run the paper stands behind
2026-09-17T08-41-25_analyze_e1qneg	the three filed accounts, scored by the filed rule

A control that could not have failed. The aligned cells were pre-registered as the specificity control for C3, on the reasoning that a generic difficulty increase would raise error in both halves of the design. With the aligned contrast at $J_A = 0.988$ the aligned cells were at

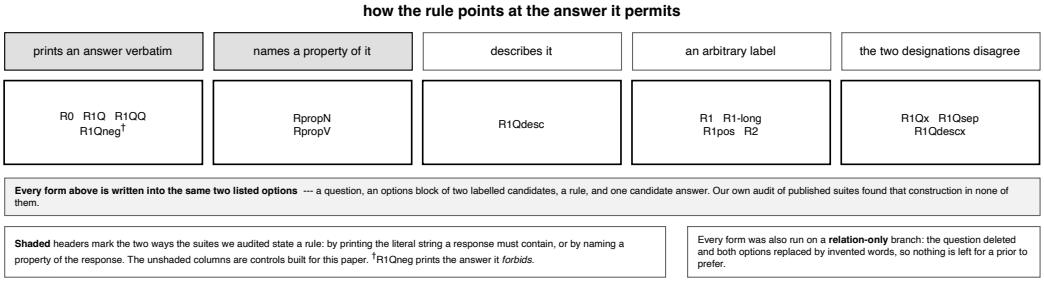

Figure 2: Every rule form this paper runs, by how the rule points at the answer it permits. Shaded headers mark the two ways the suites audited below state a rule; the unshaded columns are controls built for this paper, and the two-option block every form here sits in was found in none of the audited suites. Drawn, not measured: no quantity in this figure comes from a run.

ceiling before any model ran, so the control was bounded near zero by construction. It is reported as a failed control rather than as the control it was registered as, and the crossed R1Q cell carries the argument instead.

Analysis runs retired for provenance, not for their numbers. Two runs of the E1 quote analysis produced the same numbers as the one the paper cites and were retired because the working tree was not clean when they ran: the code behind them is not recoverable from git, which is the only property that matters here. They remain on disk with their envelopes. One generation run of the relation-only arm was voided mid-flight for the same reason and rerun; its directory carries a VOID.txt saying so, and no number in this paper descends from it.

2026-09-17T08-08-48_analyze_e1q dirty tree; superseded, same inputs and code
2026-09-17T08-18-42_analyze_e1q dirty tree (untracked file); superseded
2026-09-17T06-30-01_exp_e2q_relonly_audit VOID: dirty-tree SHA recorded, killed and rerun

An earlier study, kept beside this one. The sections, claim table and runs of a previous paper on prior strength remain in paper/sections/ and experiments/logs/. They are not part of this submission, no claim here rests on them, and the one result of theirs this paper touches — the criterion-naming effect — is corrected in Section 4.3 against a matched baseline rather than defended.

S10 Adjacent literatures, separated

tasks (Huang et al., 2024; Tyen et al., 2024; Kamoi et al., 2024; Stechly et al., 2025; Binder et al., 2025), response habits move scores (Zheng et al., 2024; Braun, 2025), and probes recover what the prompt makes salient rather than knowledge (Farquhar et al., 2023). None of them concerns what an endorsement rate identifies when the requirement is in the context window; Appendix S10 separates each from the question asked here. Nothing in this paper rests on mechanistic localisation, no activation or probe evidence is offered, and no causal language is used.

S11 The benchmark rule-form audit

The coding rule is the final section of the pre-registration reproduced in Appendix S1, fixed before any count existed. Templates were collected from primary sources by a process instructed to quote and not to classify; text that could not be quoted is recorded as absent or marked a paraphrase, and neither is counted. Of 21 recorded items across 13 suites, 16

carry a verbatim quotation, 2 a marked paraphrase, and 3 publish no rule text at all — the last is itself worth recording, since a benchmark whose requirement is never written down cannot be audited by anyone.

The primary count is IFEval's verifiable-instruction types, because that population is enumerable and published in full in instructions.py. Each of the 24 templates is coded in experiments/data/ifeval_coding.tsv with the template text and a one-line reason beside it, and the analysis refuses to run if the coding table and the source templates disagree about which types exist. A reader can check every assignment against the quoted source rather than taking the count on trust.

What the count supports is that the rule form manipulated in this paper is the form published evaluations use. It does not support any claim about those benchmarks' scores: their items are not ours, no model was run on them here, and the effect sizes we measure transfer only under an assumption we do not test.

S12 The registered specificity control in full

Q_A is bounded by how much room the aligned cells have, and on this bank they have very little. The aligned contrast is $J_A = 0.988$ under the quoting rule and 0.982 under the permissive label rule, reaching 1 in the most saturated cells, so mean aligned-cell error is 0.006 and 0.009 respectively. Two consequences follow and we state both. First, Q_A could not have been large, so this control is evidence that the cost of removing the quote is concentrated in the conflict cells, and is not evidence that a model which ignores the label would have been caught — in an aligned cell compliance and baseline correctness agree, so answering from the baseline answer alone already scores near ceiling. Second, with J_A this close to one, $R_C = \frac{1}{2}(J_A + J_X)$ is nearly an affine rescaling of J_X on these data, so the fall in R_C and the sign change in J_X should be read as one finding and not two. The prohibitive form is the only one with appreciable aligned headroom ($J_A = 0.864$, aligned error 0.068), and it is also the form §4.3 disqualifies for half the roster, so we do not lean on it. A control that can actually fail requires aligned items on which the baseline answer is unavailable; we did not run one, and §5 says what it would take.

Removing the quote adds an indirection step, so we pre-registered the aligned cells as the specificity control: a generic difficulty increase raises error in both halves of the design. Over the permissive form and the models that pass the sanity arm of Section 4.3, the conflict-cell cost is $Q_X = 0.156$ $[0.134, 0.178]$ and the aligned-cell cost is $Q_A = 0.004$ $[-0.002, 0.009]$, with $Q_X - Q_A = 0.153$ $[0.130, 0.175]$. All three pre-declared conditions hold: the roster mean above 0.10, $Q_X - Q_A$ above 0.05, and 3 of four models resolved individually.

S13 Deviations from the filed plan

Every departure from the pre-registration reproduced in Appendix S1, including the ones that cost us a claim, is recorded here. The entries distinguish an analysis-level correction (D14), a registered arm not run (D1), post-hoc analyses and screens (D3, D5), and implementation or provenance corrections (D6, D10, D12); other entries explain changed interpretations or unmet narrative consequences. The numbered references are retained so each event can be traced to the filed plan.

D1. E3 was registered and not run. The addendum registers E3 as adding checkpoints above 8B on the main wording under R0/R1/R1Q, to answer whether the effect exists at larger scale. E3 was not run: this project has no paid inference budget and no accelerator beyond one shared workstation, and the largest checkpoint that fits alongside other work on it is 8B. No result above 8B is reported and no claim above 8B is made. The one within-family comparison we do have runs the other way — Q_{quote} is smaller on Qwen3-8B than on Qwen3-4B — and §5 says so rather than leaving it to be found.

D2. A mechanism reading was withdrawn, then re-established in a narrower form. The submitted draft read the effect as mismatch rejection against the printed string. Q_{bind}

excluded zero in the direction that reading forbids, and R1Qneg moved p_{01} opposite to its prediction, so it was withdrawn. R1Qx, filed afterwards and before it was run, established the narrower statement the paper now makes: the comparison is against whatever string the rule prints, in the polarity of the slot it appears in. The sequence is in Appendix S9.

D3. The responsiveness screen is post hoc, and the criterion it screens passes on a checkpoint that barely moves. No responsiveness floor was registered. It was added on noticing that the registered 3-of-4 criterion was passing on a checkpoint that answers no to essentially every conflict item. Both counts are reported and neither is presented as the registered one. The unscreened count is 3 of four and the screened count is 2 of 3, so the registered criterion is met on the first and not on the second. Which checkpoint is at issue is no longer left to be worked out: the per-model estimates are in Table 3, the third vote is Qwen3-8B at 0.034, below the 0.05 margin this paper uses elsewhere on its own roster means, and Llama-3.2-3B reverses. A reader who counts only responsive checkpoints should read Q_{quote} as established on two of three, and the roster-mean estimate 0.153 $[0.127, 0.180]$ as carried by two checkpoints rather than by the roster. Nothing in the paper's central claim rests on this count: that claim is the override arm of Appendix S28, which holds on four of four.

D15. Addendum 11 came back indeterminate, and the property claim is narrowed as that addendum said it would be. Addendum 10's result was first written up as showing that a property rule is the most expensive rule form here, and defended with aligned-cell error. That defence is invalid — in an aligned cell the compliant candidate is the true answer, so a model answering from the facts alone scores near-zero aligned error — and it is the objection this paper itself makes in Appendix S12. Addendum 11 was filed to replace it with a competence control on the relation-only branch, with both readings and both consequences fixed before the run. Neither reading fired: balanced accuracy against the property forms is 0.526 $[0.518, 0.535]$, short of the 0.6 floor that would have established competence and just above the chance value that would have established its absence. The addendum fixed that case too: report it as indeterminate and narrow Q_{prop} to a statement conditional on property competence. Accordingly the rule-form reading of Q_{prop} is withdrawn from the abstract, §4.2 and Appendix S15; the claim that this arm refutes the self-containedness account of the decomposition is withdrawn with it. The remaining observation is near-chance property-rule accuracy without a conflicting factual question. Because the registered verdict is indeterminate, the conflict-only score on these stimuli cannot be interpreted as a clean measure of rule-form sensitivity. The title and the decomposition are unaffected: every form they rest on is one this roster applies at between 0.922 and 0.987 on the same branch.

D14. The crossed forms were computed at a 0.90 interval level that nothing registered, and have been recomputed at 0.95. Section 3 of the filed plan fixes the unit, the estimator, the resample count and the seed, and says the level is the same as the primary analysis, which is 0.95. The crossed-form scripts nonetheless used 0.90 — the more permissive of the two levels appearing in this paper, never coverage-simulated on these 40 clusters, and, until this entry, absent from the list above while far smaller departures were recorded in it. Every crossed-form analysis has been recomputed at 0.95 by replaying each run from its own stored configuration. What changed: the interval bounds widened. What did not change: every point estimate, every interval's exclusion or inclusion of zero, every per-model count, and every pre-registered reading. We report that rather than presenting the corrected level as though it had been used all along, because the reason to distrust an unregistered level is that it might have been chosen for what it yielded, and the only way to answer that is to say what the stricter one yields.

D4. The order-stratified moderator analysis was late. It was registered and initially not reported. It is reported here: Q_{quote} is 0.083 $[0.066, 0.101]$ over 38 groups when the baseline-correct answer is listed first and 0.250 $[0.211, 0.287]$ over 37 groups when it is listed second. Both exclude zero, so the effect is not an artefact of option order; it is about three times larger in the second stratum, which we report as a moderator and do not interpret. A stratum loses two or three whole construction groups because option order is assigned per

base question, which is why the primary analysis refuses these slices and a separate entry point runs them.

D5. The E2 secondary analysis was not pre-registered. The switching-margin calculation in Appendix S4 was written after the null, to say why the null occurred. It is labelled secondary in the main text.

D6. Two analysis runs and one generation run were retired for provenance. None for their numbers; Appendix S9 lists them with their run ids.

D13. R1Qdescx is uninformative as built, and its addendum described the reason backwards. Addendum 9 says the description is degenerate in the aligned condition, copying a caveat that is true of R1Qdesc. For R1Qdescx the construction makes both halves confounded, in opposite directions: the description is built from the question whose true answer is the non-designated candidate, which in the conflict cells is the question already printed above the rule — so following the description and answering that question are the same act — and in the aligned cells is an independent second retrieval, so not following it may be retrieval failure. Neither half licenses a reading and the four-cell mean of the two is an artefact. We report the form, report that no conclusion is drawn from it, and name the fix: a description matched for retrieval cost across the two halves. The filed text stands unaltered; the error in it was ours. For completeness, Q_{incons} for this form is 0.273 $[0.248, 0.298]$, and the per-half figures are -0.423 and 0.029.

D12. Addendum 3 refers to a floor that no addendum declares. Its threshold paragraph says responsiveness is reported "using the same 0.10 conflict-cell floor declared in Section 8's analysis". Section 8 declares no such floor: the 0.10 screen is post hoc, which is what D3 says. The filed text is reproduced unaltered because altering a filed document to remove an embarrassment is worse than the embarrassment, and this entry records that the cross-reference inside it is wrong. Nothing in the paper's counts depends on it — both the screened and unscreened counts are reported everywhere the screen applies.

D11. Both of addendum 8's readings fired, and its consequence is applied by qualifying the title rather than deleting a clause. That addendum fixed two thresholds and wrote them as exclusive. Both were met: the string still governs when detached ($X_{\text{desig}} = -0.356$, negative on 4 of four), and the apposition contrast 0.095 $[0.071, 0.119]$ clears its own threshold. The filed consequence of the second was that the title's third clause be withdrawn. Deleting it would misreport a result that holds on every checkpoint, so the clause is narrowed instead — it now says a printed answer overrides the rule's designation, which is what both forms show — and the exclusivity assumption in the addendum is recorded here as the error it was. The second threshold was also not computed at first: the analysis script loaded only R1Qsep and R1Q, so a registered decision went unreported until it was recomputed from the logs.

D9. The planned title change for branch B was not implemented. Section 5 of the pre-registration assigned each outcome a consequence in advance. The observed outcome — Q_{quote} established in E1 with E2 below its margin — corresponds to branch B, which called for the title to name an interaction. We did not implement that title change because the interaction was not estimated: the two arms differ in scale, sample and wording. We record this as an unmet pre-registered narrative consequence.

D10. An unplanned replication. The R1Qneg generation rebuilt R1Q, R1QQ and R1-long as well, so 7680 stimuli were run twice on the same models under the same greedy decoding. 3 answers differed, all on one checkpoint. We report it because it is the only direct evidence here about run-to-run stability, and because pooling the two copies would have doubled the denominators in Appendix S2 while leaving the rates right — which it did, until the table was deduplicated.

D8. The crossed contrast was registered on one wording and has since been run on the other two. Q_{quote} is 0.188 and 0.185 on the paired wordings against 0.153 on the bridge wording, excluding zero on 3 of four in each. This was filed as a robustness check with no threshold, and the registered claim remains the bridge-wording estimate.

D7. Three forms were added after the first analysis, each filed before it ran. R1Qneg, R1Qx and R1pos are not in the original addendum. Each was registered in a numbered addendum naming its competing accounts and their predicted signs before the form existed, and in two of the three the prediction we wrote down as the one we expected was the one that failed. The filed text of all of them is reproduced in Appendix S1, so a reader can check the predictions against the results rather than taking the order on trust.

S14 R1Qneg in full

R1Qneg quotes only the forbidden answer. Endorse whatever string the rule prints predicts p_{01} rises; it falls, 0.158 $[0.130, 0.188]$, with 0 of four in the predicted direction, while L_X improves on R1 by 0.224 on 4 of four. A printed string helps even when what it names is the answer to refuse.

S15 The property rule

Every form in this paper except the two below designates a member of the options block — by quoting it, by an arbitrary letter, by an ordinal, or by a description attached to a letter. The audit of Appendix S11 codes published verifiable instructions into three kinds, and the largest of them is the kind where compliance is decided by a property of the response, with no string in the rule to match and no options block to resolve. Until the run reported here, no form in this paper was of that kind, so the contrast the field itself faces had no estimate in it. Addendum 10, filed before the run, says so and fixes what follows from each outcome.

The two forms. rpropn states a count property, only an answer of exactly n letters is permitted, with n taken from the designated candidate; it applies only where the two candidates differ in that count, which costs it two of 40 construction groups. rpropv takes the first predicate from a list fixed in the addendum that holds of the designated candidate and fails of the other — single word, more than one word, begins or does not begin with a vowel, some or no repeated letter, first letter in the first or second half of the alphabet, and a count clause only as a last resort. It keeps all 40 groups. Neither prints a candidate, names a label, or fixes a position; each rule's property is re-derived from the rule sentence and checked to hold of the designated candidate and fail of the other before the stimulus is emitted. Each contrast is computed over the questions the forms it compares all cover.

Table 2: Conflict-cell error L_X by rule form, with aligned error L_A and e_{01}, the rate of endorsing the candidate that is correct and forbidden. Q_{prop} and Q_{propind} are as defined in addendum 10. Intervals are 0.95 stratified paired cluster percentile bootstrap.

rule form	L_X	L_A	e_{01}
R0, prints the permitted answer	0.455	0.002	0.002
R1, arbitrary label	0.554	0.002	0.166
rpropn, count property	0.939	0.052	0.916
rpropv, mixed property	0.905	0.058	0.880
Q_{prop} rpropn 0.483 $[0.462, 0.504]$, positive on 4 of four			
Q_{prop} rpropv 0.449 $[0.424, 0.473]$, positive on 4 of four			
Q_{propind} rpropn -0.385 $[-0.409, -0.360]$, rpropv -0.351 $[-0.378, -0.324]$			

What the numbers say. A property rule is the most expensive form in this paper. It costs more than a rule that prints its answer, which addendum 10 registered as the matchability

account's prediction, and it also costs more than arbitrary-label indirection, which no account predicted. The registered reading "a property costs" is established on both forms and on 4 of four checkpoints each; the registered reading "a property is free" is not.

Is it that the task is harder? The aligned cells cannot say. An earlier version of this section argued that it is not, on the grounds that aligned error under the property forms is only 0.052 and 0.058. That argument is invalid and we withdraw it. In an aligned cell the compliant candidate is the true answer, so a model that ignores the rule entirely and answers from the facts scores near-zero aligned error — which is the objection this paper itself raises against aligned-cell controls in Appendix S12. The aligned cells cannot separate "applies the property" from "answers from truth", and we should not have used them to. What the conflict cells do show, without needing that inference, is where the collapse sits: e_{01} runs 0.002 under R0, 0.166 under R1 and 0.880 under rpropv. The competence question is settled separately, on a branch with no true answer to fall back on; addendum 11 (Appendix S1) fixes the reading of each outcome, including the one that withdraws the claim.

The competence control, and what it costs the claim (addendum 11). The question the aligned cells could not answer is whether these checkpoints can decide such a property at all. The relation-only branch answers it: no question, no true answer, both options either real words or invented ones, so nothing can be answered from the facts and accuracy against the rule is a direct measure of competence.

rule form on the relation-only branch	A_{rule} (pseudoword)	(real word)
R0, prints the permitted answer	0.987	0.955
R1Q, prints it and labels it	0.955	0.951
R1, arbitrary label	0.922	0.926
rpropn, count property	0.512	0.531
rpropv, mixed property	0.538	0.560
both property forms	0.526	0.547

The property forms sit at $[0.518, 0.535]$ on the pseudoword variant with 0 of four checkpoints above the 0.6 floor the addendum filed, while the label forms on the same stimuli are between 0.922 and 0.987. Neither filed reading fires exactly: the interval lies above chance, so "they cannot evaluate it" is not established either, and addendum 11 calls that the indeterminate case and fixes its consequence — Q_{prop} is reported conditional on property competence rather than as a rule-form cost. That is what we do. The reading we would have preferred is the one this arm was built to license, and it did not license it.

On these property-rule stimuli, this roster's decisions are near chance even with no conflicting fact anywhere in the prompt. The registered competence verdict remains indeterminate: the interval excludes chance but does not reach the competence floor. A conflict-only score on these forms therefore cannot isolate rule-form sensitivity from difficulty applying the property. We report the rule-form comparison as unresolved.

Per checkpoint, and the screened count. Unlike Q_{quote}, whose roster mean is carried by two of four checkpoints (Appendix S23), Q_{prop} is close to uniform across the roster and excludes zero everywhere.

checkpoint	Q_{prop} (rpropv)	0.95 interval
Gemma-3-4B	0.553	$[0.513, 0.594]$
Llama-3.2-3B	0.411	$[0.369, 0.450]$
Qwen3-4B	0.431	$[0.389, 0.470]$
Qwen3-8B	0.401	$[0.366, 0.435]$
roster mean	0.449	$[0.424, 0.473]$

Addendum 10 asked for the screened count beside the unscreened one and the first version of this analysis omitted it. All 4 checkpoints clear the 0.10 conflict-cell floor used for Q_{quote}

— including Qwen3-8B, which does not clear it on the main bank's label forms — so the screened and unscreened counts coincide at 4 of four on rpropv and 4 of four on rpropn.

The cost does not sit on the predicates that require counting. If what the property forms measure is arithmetic rather than rule form, the cost concentrates on the letter-count clause, which is the only registered predicate that requires counting anything. It does not. The breakdown below is registered in addendum 11, §17.1, as descriptive and is reported whichever way it falls; it uses the addendum-10 logs and needed no further run.

predicate	L_X	n	predicate	L_X	n
repeated letter	0.949	176	alphabet A–M	0.900	80
final vowel	0.938	160	alphabet N–Z	0.887	80
letter count	0.937	1176	initial vowel	0.866	112
no repeated letter	0.929	184	final consonant	0.854	144
single word	1	8	initial consonant	0.821	112
multi-word	1	8			

The counting clause sits in the middle of the range, not at its top, and three predicates that require no counting at all are above it. The two word-count predicates are at 1.0 on 8 conflict observations each and carry nothing; we print them rather than drop the rows that came out inconveniently small. This rules out arithmetic as the explanation. It does not rule out property evaluation being hard in general — every predicate here is orthographic — which is what addendum 11 tests on a branch where there is no true answer to fall back on.

The alternative account this was run to test is not settled by it after all. A standing objection to the decomposition is that what makes a rule cheap is not that the compliant answer is printed but that the rule is self-contained, the expense being the return to the options block to resolve a symbol — a property of this harness that no published suite shares. A property rule looked like the case that separates the two: it requires no return to the options block, so that account predicts $Q_{\text{prop}} \approx 0$, and Q_{prop} is 0.449. We drew that conclusion and have withdrawn it. The competence control below shows the roster near chance on these properties when the rule is the only thing to go on, so a large Q_{prop} is what an unevaluable rule produces as well as what an expensive rule form produces, and this arm does not distinguish the two accounts. What would: a property these checkpoints demonstrably decide at the level they decide a label rule, and we did not find one among the predicates we registered.

S16 Identification bounds for unreadable replies

With cell total N, yes count Y and unreadable count U, the latent rate lies in $[Y/N, (Y + U)/N]$, and a linear contrast $\sum_j a_j p_j$ is bounded by taking the lower endpoint where $a_j \geq 0$ and the upper where $a_j < 0$, and the mirror for the upper bound. These are identification intervals over admissible relabellings of the unreadable replies, not confidence intervals, and we never combine the two. Filling every unreadable reply with the same answer is not a bound; it appears only as two named sensitivity fills. On the runs reported here no generation was unreadable, so every bound coincides with its point estimate.

S17 The dereference account, and the two tests against it

One account fits every form so far without mentioning strings: these models resolve an arbitrary letter poorly, and any form computable without that resolution scores better. R1pos separates it — designation by position (the answer listed second above), no candidate string, no letter to resolve — and the account predicts Q_{pos} of Q_{quote}'s order. It is negative: $Q_{\text{pos}} = -0.172$ $[-0.201, -0.144]$, on 0 of four checkpoints in the predicted direction against 3 for Q_{quote} on the same items. Removing the arbitrary letter without printing a string makes the task harder. That does not settle it by itself — R1pos substitutes an ordinal

reference for an alphabetic one rather than removing reference — so we also gate on measured competence. The relation-only arm scores label resolution directly, with no question and no printed string; restricting Q_{quote} to items on which a model answers every relation-only question correctly leaves it unchanged — 0.199 and 0.364 on the two checkpoints that carry the effect $[0.276, 0.447]$, against 0.153 ungated. On items a model provably can dereference, the printed string still buys what it buys. Post hoc; the gate is on items and comes from a different arm, and Llama-3.2-3B is competent on too few items to read (16 of 160).

S18 The C2 selection experiment in full

Does filling in the aligned cells buy a better verifier? (C2)

Twelve candidate configurations were audited by four policies at equal budget on disjoint slices: a conflict-only scalar score, a conflict-only hit/false-alarm pair, mixed coverage, and the crossed four-cell audit, with selections, scores and the hash of every visible row locked before any transfer output was scored. The workload is 400 independent rule instances labelled by a deterministic predicate before any model ran (Appendix S6).

The pre-specified comparison is the crossed audit against the full hit/false-alarm baseline: $G = -0.007$ $[-0.019, 0.004]$, which does not exclude zero and whose point estimate favours the baseline; against the scalar conflict audit, $G = -0.018$ $[-0.041, 0.006]$. No advantage to the crossed audit is established, and of the four selected configurations the one the scalar conflict audit chose has the lowest test error. That is a statement about four policy choices and not about the twelve candidates: the best configuration in the set, Qwen3-8B|W0 at 0.172, was selected by no policy. We do not claim superiority over conflict-only auditing in any form, which is the rule we fixed before running it. Both four-cell policies score a configuration by its worst cell, so the aligned half can change a selection only when aligned error rivals conflict error, and here the margin at the selected configuration is 0.643 (Appendix S4; secondary, not pre-registered). The aligned half is worth its cost when a verifier can fail on aligned items, and on this roster it essentially cannot.

S19 The relation-only arm in full

Strip the world-knowledge question away and leave two labelled options and a rule. On the permissive form all four models clear the midpoint between chance and perfect. On the prohibitive form Llama-3.2-3B is at chance, 0.494 $[0.481, 0.500]$, with 0 accuracy when the rule permits the candidate — the pattern an always-no responder gives, and this arm does not separate that from a failure to parse the negation — and Qwen3-4B reaches 0.613. 6 of 8 cells pass; numbers from a failing cell are reported as a failure to read the rule, which is why C3's primary uses the permissive form only.

S20 Limitations in full

Fixed roster, fixed templates, one domain. Four instruction-tuned checkpoints at 3–8B, two frozen paired template blocks per wording, and one bank of two-option world-knowledge verifications from a single construction pipeline. Models, templates and domain are design factors and not samples: every interval covers the sampling of new base questions and construction groups, not a population of models, prompts or tasks. With four checkpoints, one of them reversing sign, a roster mean is a dependency report and we do not offer it as more.

Scale runs against us where we can see it. Q_{quote} is smaller on Qwen3-8B than on Qwen3-4B, the only within-family comparison available here, and the registered arm above 8B was not run (Appendix S13). Nothing in this paper establishes the effect at a scale where verifiers are usually deployed.

One shortcut closed, not all of them. Label-mediated rules remove the printed string, and §4.2 shows that string was doing the work, but nothing establishes that what remains is rule application in any deeper sense. The aligned cells registered as the difficulty control could not have failed, and the crossed cells carry the argument instead.

R_C and R_T are response contrasts. Differences in endorsement frequency between conditions — not capacities, knowledge, or causal effects of truth. No activation or probe evidence is offered and no latent mechanism is identified.

Unreadable replies are bounded, not imputed. Those bounds are identification intervals over admissible relabellings, not confidence intervals, and we never combine the two (Appendix S16). On the runs reported here no generation was unreadable.

The transfer workload is synthetic. Gold labels from a deterministic predicate — trustworthy for that reason and narrow for the same one — so neither its benefit nor its null transfers to arbitrary real verification tasks.

The prohibitive arm does not separate parsing from a response habit. An always-no responder produces the same numbers as a model that cannot read a prohibitive rule, and §4.3 does not distinguish them. Either way a compliance number from such a cell is a measurement of the interface rather than of compliance.

A control that can actually fail was not run. It needs aligned items on which the baseline answer is unavailable. We did not build them, and §5 is where that stays open.

S21 Endpoint and sampling unit, in full

Three outcomes, and an endpoint that does not impute

Every generation falls in exactly one of yes, no, unreadable. Transport and runtime failures are recorded separately and are never silently converted into either — an unreadable means the model answered and the answer could not be read.

The primary endpoint is operational:

$$e_{ct} = \Pr(\text{no correct readable verification label} \mid C = c, T = t),$$

where correct is yes when C=1 and no when C=0, and unreadable counts as a failure. This scores whether the interface completed the verification task. It imputes nothing about a latent yes or no. From it, $L_A = \frac{1}{2}(e_{11} + e_{00})$, $L_X = \frac{1}{2}(e_{10} + e_{01})$.

We report the complete-case quantities p_{ct}, J_A, J_X, R_C, R_T alongside, never instead. Where unreadable occurs we give identification bounds over admissible relabellings rather than filling them with one answer, which is not a bound and is reported only as a named sensitivity fill (Appendix S16).

Sampling unit

Base questions are authored in swap groups of four from one semantic class, the false candidate of each being another member's answer under a derangement, so the four are jointly dependent and the group is the independent sampling unit. One resample moves every row of a group together: all models, cells, rule forms and wordings. Roster and template blocks are fixed design factors, so intervals cover new questions and construction groups — not a population of models or of prompts.

S22 Interval bounds for estimates the main text reports as points

Reviewers of an earlier draft found the results section too dense to read, and in particular that a headline sentence could not be traced to a named estimate without recomputing roster

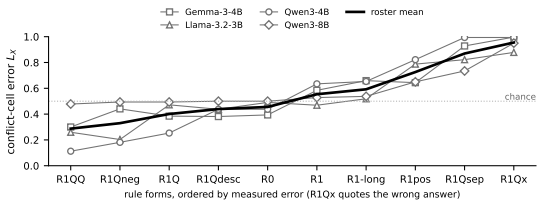

Figure 3: Conflict-cell error by rule form, ordered by the roster mean at draw time; the values are Table 1. The low-error end holds the forms whose rule writes its extension out truthfully — by quoting, by naming the forbidden option, or by describing the permitted one; the high-error end holds those that write out nothing (R1, R1-long, R1pos) and the two that write out the wrong answer (R1Qsep, R1Qx). R1pos drops the arbitrary letter while printing nothing and is worse than the forms that keep it.

means from the per-cell tables. Table 1 in the main text answers the tracing complaint. This section answers the density one without deleting anything: where the main text now gives a point estimate alone, its interval is here, at the registered 0.95 level, together with the per-wording and per-form detail moved out of the same paragraphs. Nothing in this section is new evidence and nothing was dropped from the paper to make room — every entry appeared in the main text of the previous version and appears here instead.

estimand	where the main text cites it	0.95 interval
Q_{bind}	string presence vs binding	$[0.097, 0.130]$
Q_{quote}, wording Wn	crossed forms, paired wordings	$[0.160, 0.217]$
Q_{quote}, wording Wc	crossed forms, paired wordings	$[0.156, 0.216]$
Q_{desc}	explicitness vs matchability	$[0.090, 0.139]$
$Q_{\text{desc}} - Q_{\text{quote}}$	the same paragraph	$[-0.055, -0.023]$
endorsement of the quoted option, R1Qx	the substitution structure	$[0.481, 0.512]$
quote effect, relation-only (E2)	with no prior available	$[0.017, 0.033]$

Three further quantities moved here from the same paragraphs. Under R1Q the designated candidate is endorsed at 0.598 against 8.0×10^{-4} for the other one; this is the structure that reappears for the quoted candidate under R1Qx, and the comparison is the point of that sentence. Restricted to the single wording whose per-cell table Appendix S2 reports, mean J_X under the quoting rule is 0.089 over 4 cells, against the all-wording mean the main text gives. And on the paired wording Wn, Qwen3-8B — the checkpoint that barely responds at the bridge wording — gives $Q_{\text{quote}} = 0.156$ there.

S23 The decomposition, per checkpoint

The roster means hide a split that matters for how the two components should be read.

The main text reports Q_{quote}, Q_{ind} and Q_{desc} as roster means, and the registered criterion for each is a count over checkpoints. Until now the counts were printed and the per-checkpoint estimates they count were not, which asks a reader to take the count on trust. They are below, at the registered 0.95 level.

Read across the row rather than down the column: the effect is large on two checkpoints, small on one and absent on the fourth, and the roster mean is not a statement about a typical checkpoint. The same heterogeneity does not appear on the override arm, where all four agree (Appendix S28); a checkpoint can be unresponsive on the conflict cells of this bank and decisive on that one, and Qwen3-8B is exactly that case.

Table 3: Per-checkpoint $Q_{\text{quote}} = L_X(\text{R1}) - L_X(\text{R1Q})$, with 0.95 stratified paired cluster percentile bootstrap intervals over 40 construction groups. The registered criterion is that at least three of four exclude zero in the positive direction; it is met, and the third of those three is Qwen3-8B, whose estimate is smaller than the margin this paper applies to its own roster means and whose conflict-cell endorsement is near zero throughout (Appendix S13, D3).

checkpoint	Q_{quote}	0.95 interval
Gemma-3-4B	0.200	$[0.156, 0.247]$
Llama-3.2-3B	-0.003	$[-0.028, 0.022]$
Qwen3-4B	0.381	$[0.312, 0.450]$
Qwen3-8B	0.034	$[0.016, 0.053]$
roster mean	0.153	$[0.127, 0.180]$

checkpoint	Q_{desc}	Q_{quote}
Gemma-3-4B	0.203	0.200
Llama-3.2-3B	0.031	-0.003
Qwen3-4B	0.197	0.381
Qwen3-8B	0.028	0.034

On three of the four checkpoints a description recovers the whole of the quote's effect, and on Llama-3.2-3B it recovers more than the quote itself buys. The difference between the columns — the matchability component — carries a per-model interval of its own: 0.003 $[-0.044, 0.053]$, 0.034 $[-0.006, 0.078]$, -0.184 $[-0.237, -0.131]$ and -0.006 $[-0.016, 0]$. It excludes zero on 1 of four — Qwen3-4B — against the three of four this paper requires of every other claim, and on Llama-3.2-3B it runs the other way. Under the responsiveness screen the counts are 2 of 3 for Q_{desc} and 1 of 3 for the difference, so the asymmetry between the two components is not an artefact of counting one of them on a checkpoint the screen excludes. The roster-average difference excludes zero, but a roster average over four checkpoints with one carrying the effect is a dependency report and not a population estimate. **The matchability component is reported and not claimed**, and the title names explicitness alone for that reason.

S24 Multiplicity

This paper reports on the order of sixty intervals, all of them now at a nominal 0.95 (D14). It is the C1 estimator whose coverage the pre-registration simulated (0.928–0.948, just under nominal); no coverage study exists for the crossed-form intervals, and sharing a nominal level with the C1 arm does not give them one. The missing study is not corrected for, so we say plainly which intervals carry a decision and which do not.

The decisions. Each pre-registered arm names one primary estimand and one threshold, fixed before its data existed: Q_{quote} for E1, the D rule for C1, G for C2, the balanced-accuracy margin for E2, and one estimand each for R1Qneg, R1Qx, R1pos, R1Qdesc and R1Qsep. That is ten decisions, each with a per-model count attached, and the branch each outcome selects was written down in advance. Every one of them is reported whichever way it came out, and four came out against the reading we expected.

Everything else is descriptive. The per-cell rates, the per-model tables, the wording replication, the order-stratified moderator, the dereference gating and the endorsement splits carry no threshold and select no branch. Their intervals are there so a reader can see the spread behind a mean, and they should be read that way rather than as sixty simultaneous tests.

What a correction would do to the ten. At a Bonferroni level over the ten registered decisions, Q_{quote}, Q_{desc}, X_{desig} under both R1Qx and R1Qsep, and the relation-only A_{label}

would all still clear zero by a wide margin — their bounds are far from it. The two that would not survive are the ones we already report as failures: E2 against its margin, and C2's G. No decision in this paper changes sign under correction; what changes is the confidence attached to the two negatives, and they are negatives either way.

S25 Two registered estimands reported for completeness

Addendum 2 registers $P_{\text{neg10}} = p_{10}(\text{R1}) - p_{10}(\text{R1Qneg})$ beside P_{neg01}, and addendum 3 registers $Q_{\text{incons}} = L_X(\text{R1Qx}) - L_X(\text{R1Q})$ beside X_{desig}. Neither carries a threshold and neither appears in the main text, so they are given here: $P_{\text{neg10}} = -0.291$ $[-0.325, -0.255]$ — endorsement of the designated candidate rises when the forbidden answer is quoted — and $Q_{\text{incons}} = 0.554$ $[0.531, 0.577]$, the cost of an inconsistent quotation scored against the label binding. A registered quantity that is computed and then left out of the write-up is the defect this appendix exists to prevent.

S26 R1Qsep in full

In R1Qx the contradicting string sits in an apposition on the label it contradicts, so a model reading it as a gloss on that label would be applying a coherent rule. R1Qsep moves the string into its own sentence. The apposition carries part of the effect — the registered contrast against R1Qx is 0.095 $[0.071, 0.119]$, above the 0.05 filed for it (Appendix S13) — and not most of it: the override weakens and survives — $X_{\text{desig}} = -0.356$ $[-0.381, -0.330]$ against -0.452 in apposition, negative on 4 of four checkpoints, with the label-designated candidate endorsed at 0.026. Part of the effect is carried by the apposition and most of it is not.

S27 R1Qdescx in full

The description control, and why it does not settle the comparison. R1Qdescx points a description at the other candidate. It is uninformative as built, and in a way we did not see until after running it: in the conflict cells the description restates the question already printed above the rule, so following it is indistinguishable from answering that question; in the aligned cells it names a second question, so failing to follow it may be retrieval failure rather than a preference for the label. The halves accordingly disagree (-0.423 against 0.558-scale in the other direction), and we draw nothing from either. Whether a description overrides a binding as a quotation does is open, and a retrieval-matched description would settle it (Appendix S13, D13).

S28 The arm that carries the claim, in full

This is the arm the paper says its central claim rests on, so it gets the fullest reporting rather than the shortest. R1Qx on the relation-only branch: no question, no true answer, and a variant whose two options are invented words. Accuracy is against the label binding, which is the only ground truth there.

checkpoint	A_{label} (real)	A_{label} (pseudoword)
Gemma-3-4B	0.039	0.004
Llama-3.2-3B	0.412	0.393
Qwen3-4B	0.036	0.015
Qwen3-8B	0.011	0
roster mean	0.124	0.103
under R1Q	0.951	0.955

The roster-mean intervals are $[0.115, 0.134]$ and $[0.096, 0.110]$. Addendum 6 fixed the threshold as the model-average interval lying entirely below 0.5 and at least three of four models

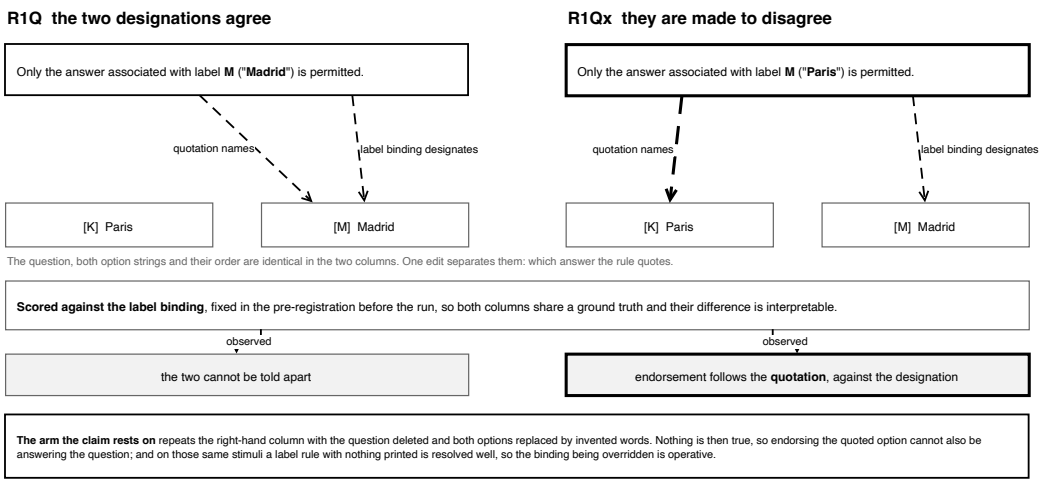

Figure 4: The override test. R1Q's label binding and its quotation point at the same option, so a model following either scores alike; R1Qx changes the quoted string alone and makes them point at different options. Scoring is against the binding, fixed in the pre-registration before the run. Drawn, not measured: no quantity in this figure comes from a run.

individually doing so, on the pseudoword variant as well as the real one. The counts are 4 and 4 of four, so the threshold is met on both variants, and 1 records the decision as the script evaluated it.

What the models do instead of following the label: on the pseudoword variant the option the rule quoted is endorsed at 0.806 and the one its label designated at 0.012; on the real-word variant, 0.791 against 0.040. The binding being overridden is operative on these same stimuli: under R1, a permissive label rule with no string printed, balanced accuracy is 0.922.

Those two splits are roster means, and a roster mean carried by one checkpoint is indistinguishable from one four checkpoints agree on unless the per-checkpoint numbers are printed. They are:

	pseudoword		real word	
checkpoint	quoted	designated	quoted	designated
Gemma-3-4B	0.992	0	0.922	0
Llama-3.2-3B	0.233	0.019	0.264	0.087
Qwen3-4B	1	0.030	1	0.072
Qwen3-8B	1	0	0.980	0.002
roster mean	0.806	0.012	0.791	0.040

The quoted option is endorsed more than the label-designated one on 4 of four checkpoints on the pseudoword variant and 4 of four on the real-word variant. Llama-3.2-3B is much the weakest of the four and points the same way. Qwen3-8B, which contributes almost nothing to the decomposition of Appendix S23, is at a ceiling here — evidence that its behaviour on the conflict cells of the main bank is a property of those cells and not a checkpoint that cannot be moved.