EVIDENCE RESPONSE IS NOT DECISION RETENTION: AN AUDIT OF AN OLFACTORY LLM INTERFACE

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a decision-level audit of predictor-backed language interfaces, separating behavioral response, assignment advantage, retention, and added value. We apply it to a fixed olfactory predictor, using released odor annotations and, separately, human applicability ratings. On 101 floral annotation pairs, two complete Qwen3-4B interfaces both make 158 correct choices across 202 presentations, yet retain 67/78 and 78/78 rule-correct pairs in both orders. A fixed vote across four label–position encodings recovers all 78 for the first interface without correcting any rule error relative to the annotations. Voting is not uniformly beneficial: on odorless pairs, it makes 18 errors among 63 accepted pairs, compared with 10 for numerical gap selection at the same coverage. In the separate human-rating task, Phi-4-mini and Qwen3-4B reproduce the score rule on every clear floral presentation under revised wording. One 32B configuration correctly reads all 160 supplied reference-rating presentations, but shows no observed matched-coverage risk advantage over numerical selectors on clear pairs when given predictions. The audit distinguishes recovery of correct decisions from improvement beyond the visible-score rule.

1 INTRODUCTION

A language interface can match a molecular predictor's accuracy while losing some of its correct decisions. Model-regression research establishes this distinction between aggregate accuracy and preservation of correct predictions (Yan et al., 2021; Xie et al., 2021). We study it with the predictor held fixed, auditing the language-model stage that receives its numerical evidence.

Olfactory predictors provide descriptor scores that a language interface can combine with molecular structure and ensemble dispersion (Lee et al., 2023b; Makri et al., 2026; Lee et al., 2025). We compare the resulting decisions with a rule using the same displayed means, separately against released annotations and human-rated applicability under the original sample preparations. Correct rule choices define retention opportunities; explicit reference-relative rule errors define correction opportunities.

We trace decisions through evidence reassignment, label and order controls, voting, wording revision, and supplied-target reading. Each comparison asks whether responding to scores, preserving correct choices, and improving on the rule coincide. Released annotations and human ratings remain separate references.

We organize established comparisons into a reusable decision-level audit (Figure 1), rather than propose new metrics. The required records link each candidate pair and reference to displayed scores, complete messages, raw outputs, and parsed physical choices. Two matched comparisons separate evidence reassignment from rule–model gains and losses; repeated presentations assess stability, and numerical selectors compare risk at equal answer coverage. On shared floral pairs and scores, identical overall correctness (158/202) conceals eleven rule-correct pairs retained by one interface but lost by the other. Voting recovers these decisions without correcting a rule error. These findings concern one frozen predictor and the tested olfactory interfaces. The workflow can be applied to other tasks; its empirical effects there remain untested.

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.

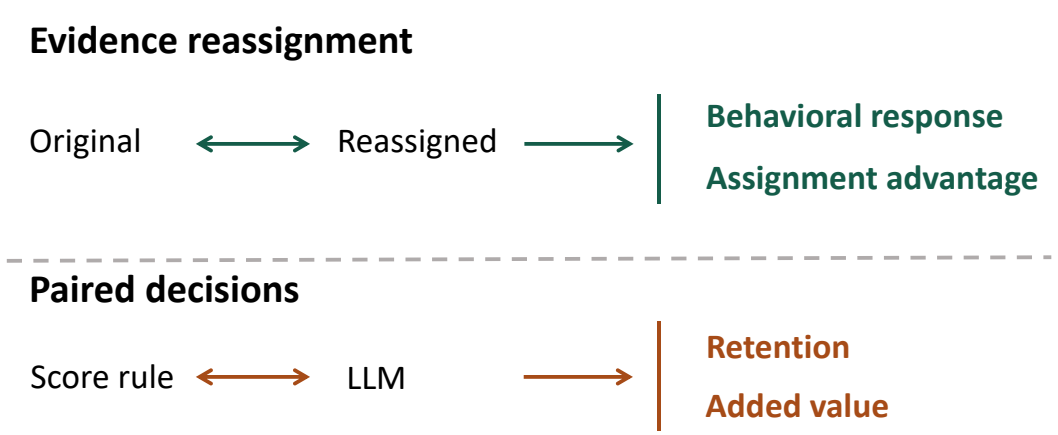

Figure 1: Two comparisons organize the audit. Reassigning evidence between fixed candidates measures behavioral response and reference-relative assignment advantage. Comparing the rule and LLM decision on each matched item measures retention and added value. Definitions and denominators are given in Section 3; the schematic contains no estimated effects.

2 Related work

Olfactory prediction and language interfaces. The principal odor map relates molecular representations to perceptual tasks and provides human measurements (Lee et al., 2023b;a). Language-model benchmarks test olfactory questions (Makri et al., 2026); electronic-nose systems combine predictions, knowledge graphs, and language generation (Lee et al., 2025). We evaluate decisions downstream of a fixed predictor, separating released annotations from human-rated applicability.

Model regression and negative flips. Model-update studies quantify negative flips despite accuracy gains in vision and NLP (Yan et al., 2021; Xie et al., 2021). We apply this distinction downstream of an unchanged predictor, combining evidence reassignment, presentation controls, and score-rule comparisons. Unlike the usual all-example negative-flip rate, retention conditions on rule-correct cases and includes abstention among losses.

Evidence response and decision robustness. Option identifiers, order, and format influence multiple-choice performance (Zheng et al., 2024; Zhao et al., 2021). Generated-text choices can be more robust than first-token readouts (Wang et al., 2024); we parse the first nonempty line and map labels to molecules. Intervention studies test whether decisions follow intermediate structures (Somov et al., 2026), while explanations can omit influential factors (Turpin et al., 2023). Here following a fallible score and matching the reference are separate outcomes.

Uncertainty and selective prediction. Risk–coverage analysis measures the trade-off between answering and error (Geifman & El-Yaniv, 2017; 2019). Related work studies unnecessary abstention in vision-language reasoning (Srinivasan et al., 2024) and uncertainty in tool-assisted models (Lymperopoulos & Sarathy, 2025). We compare models with numerical selectors at the same observed coverage. Ensemble dispersion is member disagreement, not calibrated error probability.

3 A Decision-Level Audit Framework

The audit fixes candidates, displayed scores, a numerical rule, and a reference, mapping outputs to candidate identities. For molecules x_a, x_b and query q, the frozen predictor returns a mean $\mu(x, q)$ and population standard deviation $s(x, q)$ across five members. Condition e specifies the displayed evidence. In the core audit, presentations $r \in \{0, 1\}$ exchange molecule positions; outputs are $Y_{ir}^{e} \in \{A, B, \texttt{TIE}, \text{invalid}\}$. Reference preference g_i is independent of presentation: annotated-positive versus released-zero for database labels, and fixed panel-mean ordering for human ratings. References are analyzed separately; Section 5.5 extends presentation controls to four encodings.

Behavioral response and assignment advantage. We first assess the frozen scores against the reference without an LLM. Direct evidence D and D-swap then exchange numerical blocks between molecules at fixed presentation. Map each output to a physical decision state $Z^e_{ir} \in \{x_a, x_b, \texttt{TIE}, \text{invalid}\}$. The reference-free change rate and reference-relative assignment advantage are

$$R_{\text{change}} = \frac{1}{2n}\sum_{i,r} \mathbf{1}\{Z^D_{ir} \neq Z^{D\text{-swap}}_{ir}\}, \qquad \Delta_{\text{assign}} = \frac{1}{2n}\sum_{i,r}(C^D_{ir} - C^{D\text{-swap}}_{ir}), \tag{1}$$

where C^e_{ir} indicates correctness against the stated reference. With a unique reference-preferred molecule, a nontied binary score rule reverses every choice after a swap, yet has $\Delta_{\text{assign}} = 2p - 1$, near zero when its original accuracy p is near one half. Neither estimand measures retention or added value.

Retention and order stability. Let H_{ir} denote correctness of the visible-score rule h, which compares the same three-decimal means and returns `TIE` on equality. Define $B^h_i = H_{i0}H_{i1}$ and $B^m_i = C^m_{i0}C^m_{i1}$. Presentation-level retention and stable retention are

$$R_{\text{keep}} = \frac{\sum_{i,r} H_{ir}C^m_{ir}}{\sum_{i,r} H_{ir}}, \qquad R_{\text{stable}} = \frac{\sum_i B^h_i B^m_i}{\sum_i B^h_i}. \tag{2}$$

Both condition on rule-correct cases. We also report $\sum_i B^m_i/n$ over all pairs and the fraction selecting the same molecule, right or wrong. This molecular consistency also reflects coverage, since abstention is not a molecular choice.

Stable correction and net added value. If E^h_i indicates that the rule makes an explicit wrong choice in both orders, stable correction is $\sum_i E^h_i B^m_i / \sum_i E^h_i$. Rule ties are tabulated separately, not counted as explicit errors. Net added correctness uses the full presentation denominator:

$$\Delta_{\text{rule}} = \frac{1}{2n}\sum_{i,r}(C^m_{ir} - H_{ir}) = \frac{G - L}{2n}, \tag{3}$$

where G counts rule-noncorrect presentations becoming correct and L counts rule-correct presentations becoming noncorrect. We split G into rule errors and ties, and L into wrong choices, abstentions, and invalid outputs. A rule copy has $R_{\text{keep}} = R_{\text{stable}} = 1$ and $\Delta_{\text{rule}} = 0$; equal aggregate accuracy does not imply equal retained choices. Ratios with zero denominators are undefined. All outcomes use their stated reference. Conditions with SMILES or dispersion provide information beyond the means available to h.

Selective value. For N presentations, C correct and W explicit wrong choices, coverage is $(C + W)/N$ and accepted-answer risk is $W/(C + W)$, undefined at zero coverage. Overall correctness C/N retains abstentions and invalid outputs. Numerical selectors rank by visible gap $|\mu_a - \mu_b|$ or normalized gap $|\mu_a - \mu_b|/(s_a + s_b + 10^{-3})$, with fixed label-free tie-breaking, and retain as many presentations as the model answers. Displayed ties can reduce the rule's actual answer count; we compare risks only at matching actual coverage. Selected examples can differ. Matched-coverage comparisons do not establish risk guarantees or statistical equivalence.

4 Study Design

4.1 Predictor and Reference Data

A fixed source of fallible evidence. The predictor combines 1,024-bit Morgan fingerprints (Rogers & Hahn, 2010) with frozen MiniLM descriptor embeddings (Reimers & Gurevych, 2019; Wang et al., 2020). Five independently initialized members use a local GoodScents–Leffingwell compilation (Barsainyan et al., 2023): 4,983 molecules split by connectivity into 3,587 weight-fitting, 399 affine-adjustment, and 997 development molecules. Supervision covers 110 of 138 descriptors; 28 are withheld from both fitting stages, with descriptor text available. The ensemble is frozen before LLM evaluation (Appendix B). Scores estimate descriptor association, not calibrated human intensity; dispersion is not a confidence interval.

Released-annotation cohort. Excluding the 4,983 training-compilation connectivities leaves 583 molecules in the available OdorNet table (Yang et al., 2026). The initial query, *floral*, is withheld from supervision and provides 101 positive–released-zero pairs without within-query molecule reuse. Extensions add *odorless* (101 pairs) and *nutty* (17): 219 pairs and 342 molecules overall, with cross-query reuse. Fixed hash-based pairing does not use predictor scores. Released zero denotes an annotation state, not verified sensory absence. Connectivity exclusion establishes separation from the documented fitting table; the upstream release and complete historical non-use of the local copy remain unestablished.

Human-rating cohort. We use human measurements from the version-1 principal-odor-map archive (Lee et al., 2023a), not its predictions. Structures, quality flags, overlap exclusions, and available ratings yield 161 eligible molecules. Four queries—*floral*, *roasted*, *caramellic*, and *meaty*—are chosen by availability and repeatability before evaluating new predictions; all are withheld from supervision. High–low panel-mean groups yield 40 pairs per query, involving 150 molecules. Ratings measure applicability (0–5) under original sample preparations, not intensity at a common concentration. Within-query ordering retains this proxy–target mismatch. Extreme-group pairing defines a diagnostic task rather than arbitrary or near-boundary comparisons.

A pair is *clear* when its panel-mean difference and lower 95% shared-subject bootstrap bound are positive. The bootstrap resamples 18 panelists with their actual repeated observations; other pairs remain ambiguous. Wording and reference-model follow-ups reuse 80 pairs (20/query; 111 molecules), including 51 clear pairs (18, 19, 7, 7). Membership is recoverable from the original 8B input manifest, but the exact historical subsampling key and initial information exposure remain unestablished. These follow-ups reuse inspected material, not a new blind sample (Appendix A).

4.2 Conditions, model configurations, and study sequence

The initial study compares no numerical predictions (N), routed descriptor predictions (I), reassigned routed evidence (I-swap), direct query-specific predictions (D), reassigned direct evidence (D-swap), and supplied reference labels (O), each in both orders. O is a positive control, not prediction from molecular information. Direct and routed evidence differ in content, so $D - I$ is not a clean algorithmic ablation.

The factorial study crosses SMILES and dispersion visibility with means fixed. A donor dispersion permutation preserves means; human-rating controls add no scores, block swaps, and supplied panel means. Phi-4-mini and Qwen3-4B use the full factorial; 8B uses a subset. Joint wording revision relocates cautions into condition-specific context, removes a transfer warning, labels evidence, and emphasizes the output contract. Only full-information and reference conditions are repeated, leaving revised-prompt SMILES and dispersion effects unidentified (Appendix F).

We evaluate Phi-4-mini-instruct, Qwen3-1.7B, Qwen3-4B-Instruct-2507, Qwen3-8B NF4, and hosted Qwen3-32B (Microsoft, 2025; Yang et al., 2025; Qwen Team, 2025). Availability differs by stage. Local inference is greedy with a 64-token cap and fixed parser; the hosted reference uses that cap in non-thinking mode. Precision, quantization, revision, and serving stack differ, precluding a model-size scaling analysis.

The original stages produced 24,426 formal responses. Label–position controls add 3,232 generations and reuse 808 answers: 27,658 formal generation events including repeated inputs, with diagnostics and anchors separate (Appendix G). Original assignment contrasts were protocol-designated; the initial rule comparison was post hoc. Follow-ups froze protocols after inspecting earlier results and reused samples; they are exploratory diagnostics, not independent confirmations.

4.3 Measurement and uncertainty

Analyses retain every output state, stratified by source, query, subset, prompt, and model. Paired contrasts average presentations within pairs and resample components linked by shared molecules; 20,000 percentile-bootstrap replicates give unadjusted 95% intervals conditional on the frozen predictor, pairs, and references. Model-effect intervals also condition on participant-bootstrap clear-pair selection; they do not propagate panel uncertainty, reconstruct pairs, or represent new populations. A

Table 1: Initial released floral task. Physical choice changes compare matched D and D-swap presentations without using reference correctness. Assignment advantages use the reference; intervals are paired, unadjusted percentage points. Neither condition has abstentions or invalid outputs. The rule compares the same displayed means.

	D correct	Swap correct	Δ_{assign}, pp [95% CI]	Choice changes
Qwen3-1.7B	101/202	101/202	0.00 [0.00, 0.00]	0/202
Phi-4-mini	142/202	79/202	31.19 [18.81, 43.07]	119/202
Qwen3-4B	158/202	56/202	50.50 [37.13, 63.37]	158/202
ARGMAX	156/202	46/202	54.46	202/202

single component after linking dispersion donors prevents interval estimation. Historical matched-coverage results are descriptive; voting intervals reselect the gap baseline within each bootstrap sample. Repeated queries, encodings, and reruns are not independent evidence.

5 RESULTS

5.1 UPSTREAM SCORES CARRY TARGET-SPECIFIC INFORMATION

On the initial released floral cohort, the frozen mean has AUROC 0.7808 (95% interval [0.7323, 0.8268]) over 583 molecules and average precision 0.4273 with 101 released positives. On the constructed pairs, the displayed-score rule is correct on 156/202 presentations. These results concern the released floral reference.

For the full clear human-rating cohort, the same rule produces correct/wrong/`TIE` counts of 46/22/0, 58/14/2, 16/12/2, and 24/4/2 for floral, roasted, caramellic, and meaty. These compare association scores with fixed panel-mean orderings under archived preparations. Released annotations, full human cohorts, and the reused 80-pair subset remain separate.

5.2 EVIDENCE RESPONSE AND ASSIGNMENT ADVANTAGE DIVERGE

Phi and Qwen3-4B change their physical choices on 119/202 (58.91%) and 158/202 (78.22%) matched presentations. Their reference-relative assignment advantages are 31.19 and 50.50 points (Table 1); Qwen3-4B's routed-evidence advantage is 23.76 points [11.88, 35.64]. Qwen3-1.7B always chooses position A and never changes physical choice under evidence reassignment at fixed order. No displayed mean is tied: the rule reverses all 202 choices, changing correctness from 156 to 46. Its 100% behavioral response and 54.46-point assignment advantage add no value over the rule itself. Appendix H.1 gives the matched state transitions.

5.3 EQUAL ACCURACY CONCEALS DIFFERENT STABLE RETENTION

On the same 101 released-label floral pairs, the initial direct-evidence interface (LF) and the factorial full-information interface (FF) each make 158/202 correct Qwen3-4B choices. Their shared displayed scores give 78 rule-correct pairs. LF retains 67 of these in both orders; FF retains all 78 (Figure 2). The eleven-pair difference therefore concerns the same opportunities, not different test samples. LF and FF are complete interface packages, described in Appendix K; this comparison does not isolate a sentence, evidence-field name, or runtime factor.

LF's transition ledger explains how these losses coexist with its overall count (Table 2). It retains 145/156 rule-correct presentations (92.95%), but only 67/78 rule-correct pairs in both orders (85.90%). Thirteen explicitly wrong rule choices become correct only in the forward presentation, when the positive molecule is in position A. Eleven correct choices become wrong only after reversal. None of the 23 rule-wrong pairs becomes correct in both orders.

For LF, 145 retained correct presentations plus 13 corrections give 158 correct choices. For FF, 156 retained presentations plus two corrections give the same total. Neither interface corrects a rule-wrong pair in both orders. Aggregate correctness therefore hides a different allocation of losses and one-order gains.

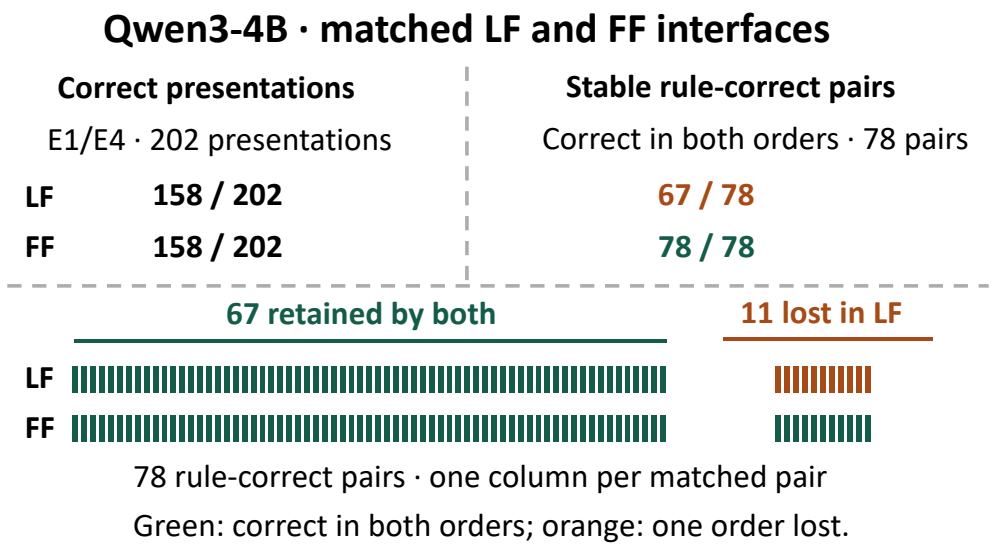

Figure 2: Equal correctness, different retention on shared molecular pairs and scores. E1/E4 are the original two orders. Each paired column represents one rule-correct pair: the highlighted eleven are lost in one LF order but retained in both FF orders. This contrasts complete interface packages, not a single prompt sentence. The supplement supplies exact pair IDs.

Table 2: Initial floral Qwen3-4B: pair-level reference-relative transitions. The positive molecule is in position A forward and B after reversal. There are no displayed rule ties or model abstentions in this comparison. All 24 unstable pairs choose position A twice.

Rule, both orders	Model, forward	Model, reversed	Pairs
Correct	Correct	Correct	67
Correct	Correct	Wrong	11
Wrong	Correct	Wrong	13
Wrong	Wrong	Wrong	10

This pattern is compatible with score following and position-A preference; it does not identify chemical correction. LF's overall both-order correctness is 67/101 versus the rule's 78/101. Appendix Table 12 reports the same accounting for all three initial models.

5.4 Net and selective gains require separate comparisons

For initial floral, the 13 favorable and 11 unfavorable transitions yield Qwen3-4B's net gain of 2/202, or 0.99 points (post-hoc 95% interval $[-3.96, 5.94]$). Phi is 6.93 points below the rule $[-12.87, -0.99]$. A local pair-selection benefit remains: among the 77 pairs where Qwen3-4B chooses the same molecule in both orders, it is correct on 67; a gap rule retaining 77 pairs is correct on 63. This descriptive advantage concerns a selected population and coexists with the retention losses above.

Across released-query extensions, consistent added value is not established. The 8B direct condition obtains 145/202 floral, 106/202 odorless, and 25/34 nutty correct presentations, compared with 156/202, 156/202, and 26/34 for the rule. Nutty has only 17 pairs. Under the original human-rating interface, Phi's clear-meaty reference-relative assignment advantage is positive (20.00 points [3.33, 33.33]), while its full-information risk is 8/24 versus 4/24 for a gap rule at matching coverage. An assignment benefit on that target likewise does not establish selective gain.

Table 3: Qwen3-4B label–position controls. LF/FF share floral pairs and scores; FO uses odorless. Correctness uses all 404 presentations; R_2 and R_4 retain rule-correct pairs in E1/E4 and all four encodings, each out of 78. Vote C/W/A counts pair decisions out of 101. Risks use matched answered counts, not matched computation. Full results appear in Appendix K.

Interface	Correct	R_2 / R_4	Vote C/W/A	Risk: vote / gap
LF	317/404	67/78; 67/78	78/20/3	20/98 / 21/98
FF	316/404	78/78; 78/78	78/21/2	21/99 / 21/99
FO	230/404	3/78; 3/78	45/18/38	18/63 / 10/63

5.5 Label and order controls separate recovery from correction

A targeted follow-up crosses molecule order with A/B assignment, giving four encodings that move whole molecule–evidence blocks together. Phi and Qwen3-4B are tested on initial floral (LF), factorial full-information floral (FF), and factorial odorless (FO), each with 101 pairs. LF/FF share pairs and scores but differ as complete interface packages; floral/odorless share 74 molecules. Numerical controls change context, retaining numbers and encoding. A fixed vote selects a physical molecule only when at least three of four outputs agree; otherwise it abstains. The vote yields one decision per pair, distinct from correctness in all four encodings (Appendix K).

The two-order contrast in Figure 2 persists across all four encodings: correctness remains close (317/404 versus 316/404), while stable retention remains 67/78 versus 78/78 (Table 3). Appendix K records the complete interfaces and their runtime alignment.

Historical and regenerated LF choices agree on all 202 shared presentations (Appendix G). For LF/Qwen, the four-encoding gain over the rule is +1.24 points [-1.49, 3.96]. Voting restores all 78 rule-correct pairs but corrects none of the 23 reference-relative rule errors, leaving overall correctness at 78/101. Each recovered pair selects (m_1, m_1, m_1, m_2) across E1–E4, with m_1 reference-correct (Appendix Table 30). Its matched-coverage risk is 20/98 versus 21/98 for gap selection: one fewer error, or -1.02 points [-2.08, 0.00]. Voting uses four model observations per pair; gap selection uses existing scores. The exploratory comparison matches coverage, not cost. FF/Qwen follows the rule on 99/101 pairs in every encoding and selects label A on the other two. On FO, voting raises risk by 12.70 points [4.55, 25.00] relative to gap selection.

Phi's numerical accuracy is 234/404 on floral and 235/404 on odorless, limiting its role as a capability-controlled replication. None of the six panels corrects a rule-wrong pair under all four encodings. Voting can nevertheless correct errors: FO/Phi corrects six and damages three within its 11 accepted pairs, obtaining 7/11 versus the rule's 4/11 on that same subset. An independent gap selector obtains 10/11 at the same coverage; Phi abstains on 90/101 pairs. These local gains do not establish a general selective benefit. The controls reveal label/position-related behavior, not its internal cause.

5.6 A complete interface revision changes retention

Information visibility under the original regime. With means present, the factorial study tests SMILES and dispersion visibility under the original prompts (Appendix E). Phi answers more often with dispersion, increasing both correct and explicit wrong answers. Qwen3-4B does not show the same pattern consistently; its average structure effect on clear roasted correctness is -10.14 points. True-versus-permuted dispersion comparisons do not establish a consistent benefit; one-component donor-linked comparisons have undefined intervals. Post-revision effects remain unidentified because only full-information and reference conditions were repeated.

A joint wording intervention. The original system distinguished predictions from human ratings and member dispersion from calibrated risk. Revision changes the placement and salience of those cautions together with the transfer warning and output instructions (Appendix F). The matched contrast estimates the package effect, not a phrase-specific effect or an identified misunderstanding.

On the fixed clear floral subset, both Phi and Qwen3-4B improve from 18/36 to 28/36 correct responses (+27.78 points [8.33, 44.44]), reaching the score rule (Figure 3). Both net increases include

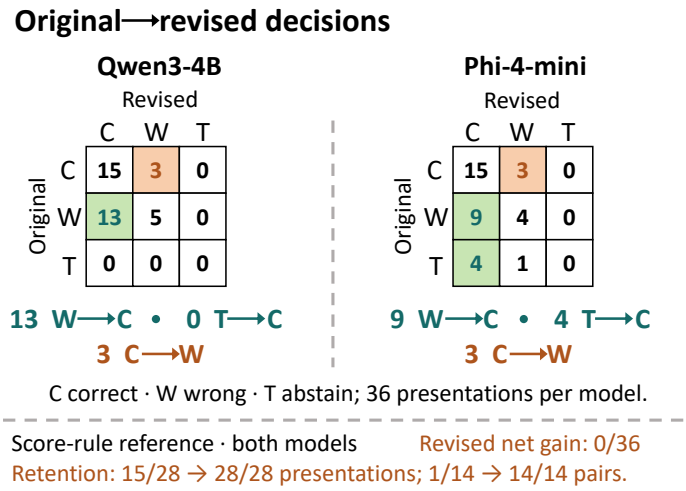

Figure 3: Joint wording revision on 18 clear human-floral pairs. Matrices separate wrong-to-correct gains from abstention-to-correct gains; both models lose three correct choices. The lower comparison instead uses the score rule as its reference: revised choices restore retention without added correctness. Exact wording is in Appendix F.

three correct-to-wrong transitions. Qwen3-4B gains 13 wrong-to-correct answers; Phi gains nine wrong-to-correct and four abstention-to-correct answers, while one abstention becomes wrong. Appendix Tables 17–20 report the full matrices.

Complete retention without added value. Under revised wording, both models agree with the rule on all 36 floral presentations. Each retains 28/28 rule-correct presentations and 14/14 rule-correct pairs in both orders, compared with 15/28 and 1/14 under the original wording. Neither corrects any of the four rule-wrong pairs in both orders; net gain over the rule is zero. Complete retention means preserving its correct choices, not solving all 18 pairs. Appendix J gives the same accounting across queries and configurations, including incomplete retention.

The change is not uniformly favorable. The 8B reference condition improves on roasted (0/38 to 11/38) and meaty (2/14 to 7/14), but falls on floral (23/36 to 17/36), adding five explicit errors; its clear predictive condition remains all-abstaining. Qwen3-4B's revised clear caramellic result exceeds the rule by one correct presentation (7 versus 6), with two rule errors corrected and one correct choice lost. This involves seven pairs; its prompt-improvement interval crosses zero. Ambiguous caramellic correctness falls from 12/26 to 11/26. Appendix I includes all query/model cells and all 29 ambiguous pairs.

5.7 Target-number reading does not establish selective value

Hosted Qwen3-32B uses the inspected 80-pair subset and revised prompt (Table 4). The positive control supplies human target numbers and evaluates selection of the larger target: 160/160 correct presentations, 80/80 both-order correct pairs, and 16 passed numerical diagnostics. This tests target-number reading, not prediction from molecular evidence or certainty about ambiguous perceptual differences.

The predictive input instead contains fallible association scores; evaluation uses the withheld panel ordering. The model answers 41/160 presentations and abstains on 119/160: 37/160 are correct and 4/160 wrong (37/41 and 4/41 among answers). All answers agree with the score rule. These counts include ambiguous pairs; they are not a controlled accuracy comparison with supplied-target reading. On clear pairs, it corrects no rule error and converts no correct rule answer into an explicit wrong choice, but turns 49 correct answers into abstentions. Successful reference reading excludes inability to compare these target numbers, but leaves proxy–target mismatch, fallible-evidence handling, wording-induced caution, and rejection thresholds unresolved.

Table 4: 32B outcomes. (a) Supplied-target reading and prediction from fallible scores are different tasks, both using all 80 pairs. (b) Predictive risk on the 51 clear pairs: N presentations, K answers. Selectors match K, potentially on different items. Risk is wrong/K, undefined when $K = 0$. No outputs are invalid.

(a) Full subset: outcome denominators are all 160 presentations

Input	Answered	Correct	Wrong	Abstained
Human target numbers	160/160	160/160	0/160	0/160
Predictor scores	41/160	37/160	4/160	119/160

(b) Clear subset: matched-coverage risk, wrong/K

Query	Answered K/N	Model	Gap	Normalized gap
Floral	22/36	4/22	4/22	4/22
Roasted	5/38	0/5	0/5	0/5
Caramellic	0/14	NA	NA	NA
Meaty	6/14	0/6	0/6	0/6

On the 51 clear pairs, Table 4 compares the model with both numerical selectors at the same actual answer count. Their observed risks coincide for each query; this is descriptive equality on the selected cases. Both-order molecular consistency is 8/18, 0/19, 0/7, and 1/7; abstentions remain in the pair denominators, so these are not pure order-error rates. Roasted's five accepted presentations and meaty's six each involve only five pairs; zero errors establish neither equivalence nor risk guarantees.

The hosted model's exact revision and precision are undisclosed. This within-configuration contrast separates target-number reading from decisions based on fallible scores; it provides neither a general capability upper bound nor a controlled model-size comparison.

6 Implications and Scope

Implications for predictor-backed decisions. Recovery and added value require different comparisons. Voting on initial floral and revised wording on human floral restore retention without stable rule-error correction. FO/Phi improves within its chosen subset but trails gap selection at matched coverage. The audit therefore reports gains, losses, and numerical selectors alongside accuracy.

For reuse, fix the reference and displayed precision, map outputs to physical candidates, and report losses alongside corrections. Separate evidence reassignment from label/order changes that preserve candidate–evidence associations. Match actual answer coverage for selective comparisons, reporting repeated-call cost separately. Appendix L specifies the records and analysis sequence.

What remains unexplained. Label/position patterns do not identify an internal cause. Phi's numerical failures limit cross-family interpretation; proxy–target mismatch, fallible-evidence handling, and instruction salience remain unresolved. The joint revision and 32B control characterize outcomes. Independent SMILES and dispersion effects were measured only under original wording.

Breadth and uncertainty. We test one predictor family, five unevenly covered configurations from two model families, and small diagnostic cohorts. Panel preparation, extreme-group pairing, and unequal repeats limit generalization. Connectivity exclusions establish neither scaffold novelty nor absence from pretraining. Source and deployment records remain incomplete. Intervals exclude adaptive prompt selection, retraining, and new populations.

7 Conclusion

The audit separates behavioral response, assignment advantage, retention, and added value using matched inputs and decision records. In these olfactory interfaces, revisions and voting recover correct choices without consistently adding value beyond the rule. Local gains coexist with losses and higher selective risk elsewhere. Independent predictors and tasks are needed to test generality.

AI USE STATEMENT

AI tools were used only for language editing; all scientific content is the authors' own.

ETHICS STATEMENT

This work reuses released molecular annotations and an existing human-rating archive; it does not recruit participants or collect new human-subject measurements. Ratings are analyzed at the level required for the archived repeated-measurement structure, without attempting to identify panelists. Molecular structures and descriptors can support beneficial scientific interfaces, but the present results do not validate safety-critical chemical, medical, or sensory decisions. Abstention is not interpreted as a certified safety mechanism. We distinguish released annotation zeros from verified sensory absence and avoid presenting model predictions as human observations. Use and redistribution of source data remain subject to their original terms.

REPRODUCIBILITY STATEMENT

The anonymous supplement reparses saved responses and reconstructs displayed-score decisions for the initial released-label floral audit, the four-encoding controls, matched original/revised human interfaces, and the 32B reference/prediction conditions. These records support the point estimates in Tables 1–4 and Figures 2–3; Figure 1 is conceptual. Full-cohort factorial results outside the included replay remain frozen aggregate exports, not independently replayed raw-response results. Historical confidence intervals and participant-bootstrap eligibility are also retained, not resampled. Appendix L and the supplement's map identify these coverage boundaries. No model calls or predictor retraining are required. Missing upstream release history and the hosted deployment's exact revision and precision remain unresolved.

REFERENCES

Aryan Amit Barsainyan, Ritesh Kumar, Pinaki Saha, and Michael Schmuker. OpenPOM: Open principal odor map. Software and data repository, 2023. URL https://github.com/ARY2260/openpom. Public implementation and associated curated odor-data resource.

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

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Brian K. Lee, Emily J. Mayhew, Benjamin Sanchez-Lengeling, Jennifer N. Wei, Wesley W. Qian, Kelsie Little, Matthew Andres, Britney B. Nguyen, Theresa Moloy, Jane K. Parker, Richard C. Gerkin, Joel D. Mainland, and Alexander B. Wiltschko. A principal odor map unifies diverse tasks in human olfactory perception. Zenodo, version 1, 2023a. URL https://zenodo.org/records/7992168. Data and code archive.

Brian K. Lee, Emily J. Mayhew, Benjamin Sanchez-Lengeling, Jennifer N. Wei, Wesley W. Qian, Kelsie A. Little, Matthew Andres, Britney B. Nguyen, Theresa Moloy, Jacob Yasonik, Jane K. Parker, Richard C. Gerkin, Joel D. Mainland, and Alexander B. Wiltschko. A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381(6661):999–1006, 2023b. doi: 10.1126/science.ade4401. URL https://doi.org/10.1126/science.ade4401.

Chung-Hong Lee, Hsin-Chang Yang, Jun-Teng Sun, and Zhen-Xin Fu. Enhancing artificial olfactory reasoning via integration of electronic nose sensing, large language models, and knowledge graphs: A case study on coffee E-Nose. *Sensors and Materials*, 37(11):4819–4841, 2025. doi: 10.

18494/SAM5826. URL https://sensors.myu-group.co.jp/article.php?ss=5826.

Panagiotis Lymperopoulos and Vasanth Sarathy. Tools in the loop: Quantifying uncertainty of LLM question answering systems that use tools, 2025. URL https://arxiv.org/abs/2505.16113. Preprint.

Eftychia Makri, Nikolaos Nakis, Laura Sisson, Gigi Minsky, Leandros Tassiulas, Vahid Satarifard, and Nicholas A. Christakis. Benchmark for assessing olfactory perception of large language models, 2026. URL https://arxiv.org/abs/2604.00002. Preprint.

Microsoft. Phi-4-mini-instruct: Official model card. Hugging Face model repository, 2025. URL https://huggingface.co/microsoft/Phi-4-mini-instruct.

Qwen Team. Qwen3-4B-Instruct-2507: Official model card. Hugging Face model repository, 2025. URL https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

Oleg Somov, Mikhail Chaichuk, Gleb Ershov, Karim Vafin, Mikhail Seleznyov, Alexander Panchenko, and Elena Tutubalina. Breaking the chain: A causal analysis of LLM faithfulness to intermediate structures, 2026. URL https://arxiv.org/abs/2603.16475v2. Preprint, version 2.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12935–12948. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.767. URL https://aclanthology.org/2024.findings-acl.767/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=qHdSA85GyZ.

Yuqing Xie, Yi-An Lai, Yuanjun Xiong, Yi Zhang, and Stefano Soatto. Regression bugs are in your model! measuring, reducing and analyzing regressions in NLP model updates. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6589–6602. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.515. URL https://aclanthology.org/2021.acl-long.515/.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14299–14308, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yan_Positive-Congruent_Training_Towards_Regression-Free_Model_Updates_CVPR_2021_paper.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388. Technical report.

Liyuan Yang, Zhiyang Liu, Bin Shao, Feng Liu, Zhuo Ma, Chunning Zhao, Xiang Wan, and Weichao Wang. SEA-driven OdorNet: A large-scale standardized molecular olfactory dataset based on a hybrid semantic taxonomy. Dataset repository, 2026. URL https://github.com/NKU-DOIE/OdorNet. Repository citation; exact historical data-release provenance is discussed in the paper.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

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*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

A Data Construction and Measurement Boundaries

Released-label tasks. The local GoodScents–Leffingwell compilation contains 4,983 molecules and 138 descriptor columns. The predictor uses 110 columns for supervision; 28 descriptor columns are withheld from both weight fitting and affine fitting. The reused development tasks comprise 75 pairs, 137 distinct connectivity identities, and 28 queries. Shared molecules connect these pairs into 62 components. We retain this cohort as development evidence, not as an untouched evaluation set.

Removing all GS connectivity identities from the local OdorNet test artifact leaves 583 rows. For *floral*, all 101 released positives are paired with 101 of 482 released zeros. For *odorless*, 101 examples are selected from each of its 307 positives and 276 zeros. For *nutty*, all 17 positives are paired with 17 of 566 zeros. Selection and pairing use fixed hash orderings rather than predictions. There is no molecule reuse within a query, but the 219 pairs across queries contain only 342 distinct connectivity identities; 91 molecules appear in multiple queries. Floral is withheld from predictor supervision, whereas odorless and nutty are supervised descriptors. A released zero denotes the absence of a database association, not a verified absence of the perceptual property. The exact upstream release binding of the local OdorNet artifact and its complete historical non-use remain unverified. The archived local input is `odornet_test.csv`; its SHA-256 is `49661f0126916207c2178772d961739b51e6db1e4d268c3b6d720a59811c0f1a`. The retained source manifest identifies this local artifact, but records an upstream main-branch URL rather than a pinned release or commit. This local identity check does not establish the missing release history.

Human measurements. We use the version-1 data archive accompanying the principal odor map study (Lee et al., 2023b;a). Individual human ratings come from `Data S3`; metadata and publicly released nominal structures come from `Data S1` and `Data S7`. Records are joined by sample identifier, not by row position. The archive's model predictions in `Data S4/S5` are not treated as human measurements. Starting from 400 metadata records, 209 have public nominal structures and 164 pass the archived author quality filters. We then exclude one GS connectivity overlap, one overlap with the released-label cohort, and one candidate without actual ratings. The resulting 161 candidates contain 5,346 observed rating rows from 18 participants. Each molecule has 30–36 rows from 15–18 participants.

The response is rate-all-that-apply descriptor applicability: selected descriptors receive scores from 1 to 5, with zero for an unselected descriptor. The reference value is the mean over the actual observed participant–repeat rows, retaining missingness and unequal repeat counts. It is not an equal-participant mean or a concentration-standardized intensity. Preparation conditions differ across molecules. We therefore interpret these values under the original measurement conditions rather than as invariant molecular properties. Comparing within-query orderings avoids requiring a common numerical scale, but it does not resolve the construct mismatch between database association and applicability under specific preparation conditions. Concentration, participant composition, and repeat weighting can also change an ordering. The visible-score rule's performance against the frozen panel means therefore measures empirical transfer to this particular target. Disagreement with those means is not, by itself, evidence that the interface lost a previously validated sensory property.

Queries, pairing, and clear-reference eligibility. Before predicting this cohort, we require a descriptor to have a panel mean for every candidate, a nonzero mean for at least 80 candidates, and between-repeat panel-mean Pearson correlation of at least 0.5. We retain floral as an anchor and select the three remaining eligible withheld descriptors with the highest repeat consistency: roasted, caramellic, and meaty. All four queries are withheld from predictor supervision; this does not imply that their broader semantic families are unseen. For each query, the 40 lowest-mean and 40 highest-mean molecules are selected, with fixed hash tie-breaking. The two groups are independently hash-ordered and paired, yielding 40 pairs without within-query molecule reuse. All 160 pairs are retained, involving 150 distinct molecules. This construction defines an extreme-group diagnostic distribution: high- versus low-mean molecules under the observed panel. Accuracy, order stability, and selective risk refer to these fixed pairs, rather than to arbitrary candidate pairs or near-boundary decisions. The construction was not chosen using predictor accuracy, but selecting the extremes of the observed target actively increases reference separation.

For a pair, let δ be the higher-minus-lower panel-mean difference. We call the reference *clear* if $\delta > 0$ and the 2.5th percentile of its participant-bootstrap distribution is strictly positive; otherwise the pair is *ambiguous*. The 5,000 bootstrap replicates use one shared draw plan over the 18 global participants, carrying each selected participant's actual repeat observations, with seed 20260918. This rule depends on human measurements, not on predictor ensemble dispersion. The full cohort has 34, 37, 15, and 15 clear pairs for the four queries, respectively.

Origin of the 20-of-40 follow-up subset. The follow-ups use exactly the original Qwen3-8B subset, with 20 of the 40 human pairs per query. The archived original protocol describes a fixed hash subset and records a freeze time of 18 September 2026, 15:00:45 UTC. Its recorded hashes match the saved pair and input files. Membership can be reconstructed by collecting distinct human-task pair IDs assigned to `qwen3_8b_nf4` in that input manifest and joining to the parent pair file. The supplement's `human_audit/provenance/subset_selection.csv` lists all 160 parent IDs and their inclusion flags; `subset_selected_ids.json` in that directory lists the 80 selected IDs. Table 5 gives the corresponding within-query indices. These records reconstruct the used sample, not the earlier sampling algorithm: the original hash key, salt, and first-selection record have not been recovered.

Human panel ratings were used to form the parent high–low groups, and human qualification statistics had been inspected. The frozen input file already contains predictor values, but this does not establish when those values were first viewed. We cannot establish whether predictor scores or earlier LLM outcomes had been inspected when the original 20-of-40 reduction was chosen. The prompt follow-up copies the prior Qwen3-8B membership and pair objects; its execution protocol was frozen at 16:36:25 UTC on the same day, after earlier labels and results had been viewed. The 32B input-transfer record at 17:13:27 UTC preserves that same membership. Neither follow-up is a new blind test.

The subset contains 111 distinct molecules and 51 clear pairs, distributed as 18, 19, 7, and 7. The remaining 29 ambiguous pairs (2, 1, 13, and 13) remain in full-subset results and are reported separately in the archived outcomes. Connectivity exclusions remove stereochemical information before canonicalization; they establish neither scaffold separation nor absence from language-model pretraining.

Table 5: Archived follow-up membership. Indices are the zero-based pair numbers in the original query-specific IDs, not a newly generated ordering. Exact full IDs and all 160 inclusion flags are supplied electronically. Each row selects 20 of 40 pairs.

Query	Selected original pair indices
Floral	2, 5, 6, 9, 11, 12, 15, 16, 21, 22, 23, 24, 26, 28, 30, 32, 33, 34, 37, 38
Roasted	1, 3, 4, 5, 8, 9, 11, 13, 17, 20, 21, 22, 23, 24, 29, 31, 32, 36, 38, 39
Caramellic	0, 2, 5, 8, 9, 12, 13, 14, 16, 18, 19, 21, 23, 24, 25, 26, 31, 32, 35, 39
Meaty	1, 2, 4, 5, 6, 8, 11, 12, 14, 15, 19, 20, 23, 25, 30, 31, 32, 35, 37, 39

B Frozen Predictor, Interfaces, and Execution

Predictor. Each of five independently initialized members receives a 1,024-bit, radius-two Morgan fingerprint (Rogers & Hahn, 2010) without feature scaling, and a frozen, normalized 384-dimensional MiniLM descriptor embedding (Reimers & Gurevych, 2019; Wang et al., 2020). The descriptor input is `olfactory odor descriptor:` {`name`}. Separate linear, LayerNorm, and GELU branches map both inputs to 128 dimensions; the molecular branch also has dropout 0.05. A 128-unit GELU scoring network consumes $[h_m, h_d, h_m \odot h_d, |h_m - h_d|]$ and adds a descriptor-specific bias. An independently fitted temperature and offset map the raw member logit z_k to a score:

$$p_k = \sigma\big((z_k + o_k)/T_k\big), \qquad \mu = \frac{1}{5}\sum_{k=1}^{5} p_k, \qquad s = \sqrt{\frac{1}{5}\sum_{k=1}^{5}(p_k - \mu)^2}. \tag{4}$$

Thus s is population member dispersion, not a calibrated error probability or a confidence interval for the human rating.

The connectivity-separated partitions contain 3,587 weight-fitting, 399 affine-fitting, and 997 development molecules. Only the 110 supervised descriptor columns enter either fit; withheld and development labels are masked. All descriptor texts remain available. Members use seeds 20260913–20260917, five epochs, at most 260,000 molecule–descriptor cells per epoch, batch size 8,192, AdamW learning rate 0.002, weight decay 0.0002, and gradient clipping at 5. The existing objective combines weighted asymmetric focal loss (positive/negative exponents $0.5/2$) with a descriptor-neighbor hinge-ranking loss of weight 0.20 and margin 0.20. Finite zeros enter as binary training targets; this does not establish an unbiased positive–unlabeled objective. There is no early stopping or development-selected checkpoint. Affine fitting uses LBFGS with learning rate 0.2 and at most 40 iterations; T_k and o_k are subsequently clamped to $[0.25, 8]$ and $[-8, 8]$. Saved member states, affine parameters, embeddings, and column order are archived. A new-process reload check compared fixed probe predictions before released-label scoring.

Evidence conditions. The original released-label experiments use six conditions. **N** contains molecular SMILES and no predictive evidence. **I** supplies routed supervised-descriptor names, means, and member dispersions. The fixed router uses two three-entry score orderings, descriptor-diversity MMR coefficient 0.82, variance scale 0.02, and soft margin floor 0.08; duplicates are removed. It does not assert that a predicted descriptor is present or absent. The queried descriptor is excluded from the routing pool in the extension. **D** supplies the query's mean and dispersion directly. **I-swap** and **D-swap** exchange the complete evidence blocks between the two molecules. **O** provides the actual released reference label. D and I expose different information and are not a same-information algorithm ablation.

The later factorial design holds predicted means fixed and varies whether SMILES and dispersion are shown, yielding four cells denoted `ID0_SD0`, `ID1_SD0`, `ID0_SD1`, and `ID1_SD1`. `SDperm` retains each molecule's mean but supplies dispersion from the next molecule in a fixed source/query hash ordering. It is a cyclic donor permutation, not merely an exchange within each pair. The human-rating cohort additionally includes N, a complete prediction-block swap, and O with actual panel means. Predicted means and dispersions are shown to three decimals; reference panel means are shown to six.

The original human-rating system text already distinguished predicted associations from human ratings and stated that ensemble dispersion was uncalibrated. The joint wording revision relocates these

cautions into condition-specific context, adds explicit evidence-type labels, removes the transfer-uncertainty sentence previously used even with reference scores, and adds a system-level instruction to follow the output contract. Molecules, numerical values, orderings, the output contract itself, and model settings are unchanged. Only the full-information and reference conditions were repeated. The comparison identifies the effect of this complete wording package; it identifies neither a particular sentence's effect nor revised-prompt SMILES or dispersion effects. Exact excerpts appear in Appendix F.

Models and parsing. Local configurations are Qwen3-1.7B in fp16, Phi-4-mini-instruct in bf16, Qwen3-4B-Instruct-2507 in bf16, and Qwen3-8B in four-bit NF4 with double quantization and bf16 computation (Yang et al., 2025; Microsoft, 2025; Qwen Team, 2025). Pinned revisions and official chat templates are retained with the inputs. Generation is greedy, one beam, batch size one, with a 64-token output budget. Thinking is disabled for the applicable Qwen3 base configurations; the 4B Instruct-2507 model uses its non-thinking template. The 32B reference uses the SiliconFlow `Qwen/Qwen3-32B` service with `enable_thinking=false`, temperature zero, top-$p = 1$, and `max_tokens=64`; no seed is sent. Its exact revision, deployment precision, and server-side template are unavailable. All 336 returned completions have stop termination and zero reported reasoning tokens. These heterogeneous deployments do not permit a parameter-count-only interpretation of differences.

The fixed parser inspects only the first nonempty output line. After trimming and case normalization, it accepts exactly A, B, or TIE with at most one permitted final punctuation mark. It does not recover an answer from subsequent explanation lines. TIE is a legal non-selection, not evidence of identical molecular properties. Invalid outputs are retained in denominators. The experiment inventory in Table 9 counts generated records rather than independent observations.

C Statistical Units and Selective Comparisons

This appendix describes the original two-order experiments. The four-encoding follow-up and its matched-coverage intervals are specified in Appendix K.

For N presentations, we separately count correct selections C, explicit wrong A/B selections W, TIE, and invalid outputs. Full-denominator accuracy is C/N, A/B coverage is $(C+W)/N$, and accepted risk is $W/(C+W)$, undefined at zero coverage. Both-directions-correct and same-physical-molecule selection use the original pair denominator. Two TIE responses do not constitute a same-molecule choice.

For a contrast, correctness differences are first averaged across the two presentations of each pair. Pairs sharing canonical molecule identities form connected components. We sample these components with replacement, retain every pair in each sampled component with its sampled multiplicity, and divide the summed pair differences by the number of sampled pairs. The archived intervals use 20,000 paired replicates and the 2.5th/97.5th linear percentiles. The initial experiment and its post-hoc argmax contrast use seed 20260916; the other stages in the original five-stage study use 20260918. The label–position follow-up uses seed 20260920 and retains all four encodings in each draw (Appendix K). Models, queries, sources, prompt versions, and truth subsets are not pooled. For dispersion-donor comparisons, donor identities also connect components. Intervals with fewer than two components are undefined. The participant bootstrap used to qualify clear human references is a separate procedure. The language-model intervals condition on the observed panel means, selected extreme groups, fixed pairings, and resulting eligibility decisions. They do not propagate uncertainty from a new panel, repeat the extreme-group selection, or reconstruct pairs in each replicate. The participant bootstrap describes uncertainty in the observed reference differences; it is not an end-to-end uncertainty analysis of data construction and model evaluation.

Intervals are unadjusted for multiple comparisons. The original D minus D-swap contrasts were designated primary; the original D minus argmax intervals were added post hoc. Follow-up protocols were fixed within each stage after earlier results had been observed. In particular, the wording revision was motivated by observed behavior, and the 32B reference used the same previously inspected subset. They are exploratory follow-ups, not independent confirmatory replications. Freezing a

Table 6: Matched prompt revision on clear human-rating pairs. Entries are *correct / explicit wrong / abstaining*; every row sums to the presentation count. Displayed means and reference pairs are unchanged. The local Qwen3-4B caramellic advantage is retained, despite its small sample. The 8B configuration abstains on all these predictive presentations under both prompts.

Query (presentations)	Model	Original prompt	Revised prompt	ARGMAX
Floral (36)	Phi	18/13/5	28/8/0	28/8/0
	Qwen3-4B	18/18/0	28/8/0	28/8/0
Roasted (38)	Phi	18/12/8	29/9/0	32/6/0
	Qwen3-4B	19/19/0	23/15/0	32/6/0
Caramellic (14)	Phi	3/3/8	6/6/2	6/6/2
	Qwen3-4B	6/6/2	7/5/2	6/6/2
Meaty (14)	Phi	6/5/3	12/0/2	12/0/2
	Qwen3-4B	6/6/2	9/3/2	12/0/2

stage limits changes during its execution; it does not make the complete sequence a single prospectively registered experiment. An interval spanning zero is not an equivalence result, and a degenerate empirical interval does not establish absence of population uncertainty.

The visible-score rule selects the larger displayed mean and returns TIE when displayed means are equal. For presentation-level coverage matching, let K be the model's observed A/B count in a source/query/truth subset. We rank presentations using either

$$g = |\mu_A - \mu_B| \qquad \text{or} \qquad g_s = \frac{|\mu_A - \mu_B|}{s_A + s_B + 0.001}, \tag{5}$$

with label-independent hash tie-breaking, retain the top K, and apply the visible-score rule. A numerical tie may reduce the rule's actual A/B count below K; such cases are not reported as equal-coverage risk comparisons. The dispersion-based rule is a same-information comparator only when dispersion is visible to the language model. Model and rule may retain different presentations. These comparisons are descriptive, with no added risk interval or guaranteed error control. On the clear 32B subset, the query denominators are 18, 19, 7, and 7 original pairs. Its five accepted roasted presentations come from five pairs; the six accepted meaty presentations come from five pairs. Equal observed risks on such selections do not establish equivalence. A paired bootstrap of a risk difference would need to recompute the model's K and reselect the rule's top K in every replicate, rather than resample only the accepted answers. We do not add such an interval here. The earlier extension instead retained an original pair only when the model selected the same physical molecule in both directions, matching the number of retained *pairs* for its rule comparison. This pair-level analysis must not be combined with the later presentation-level analysis as one risk–coverage curve.

D MATCHED PROMPT AND REFERENCE COUNTS

Table 6 gives matched wording counts across the 51 clear pairs; Figure 3 highlights the 18 clear floral pairs. Table 7 gives the clear-pair risks summarized in Table 4(b). The supplied-target and predictive counts in Table 4(a) use all 80 pairs, including the 29 ambiguous pairs; their inputs and evaluation tasks differ. Correctly reading all 160 displayed reference comparisons establishes performance on this direct target-number task. It does not establish reliable transfer from an association proxy to panel applicability, integration of uncertain predictions, or an appropriate rejection threshold. These remain possible explanations for abstention or errors with predictive evidence.

E FULL-COHORT FACTORIAL COUNTS

Table 8 reports the original-prompt factorial results for the full clear human-rating cohort. Means remain present in all four cells. Writing a_{uv} for full-denominator accuracy when structure visibility is u and dispersion visibility is v, the archived main effects are

$$\Delta_{\text{ID}} = \tfrac{1}{2}[(a_{10} - a_{00}) + (a_{11} - a_{01})], \qquad \Delta_{\text{SD}} = \tfrac{1}{2}[(a_{01} - a_{00}) + (a_{11} - a_{10})]. \tag{6}$$

The same contrasts are computed for explicit error and coverage. These effects concern visibility under the original prompt. The revised-prompt experiment does not repeat all four cells, so it cannot

Table 7: 32B reference configuration on clear human-rating pairs under the revised prompt. The model reads the displayed panel means correctly on every presentation (not shown). With predictions, its observed accepted-answer risk equals that of each numerical selector at matching coverage. These equal point estimates are not evidence of statistical equivalence; selected examples differ.

Query	Correct/wrong/TIE	Coverage	Model risk	Matched rule risk
Floral	18/4/14	22/36	4/22	4/22
Roasted	5/0/33	5/38	0/5	0/5
Caramellic	0/0/14	0/14	NA	NA
Meaty	6/0/8	6/14	0/6	0/6

Table 8: Original-prompt human-rating factorial results. Entries are correct / explicit wrong / TIE presentations; invalid counts are zero. The two condition digits indicate SMILES and dispersion visibility, respectively. N counts both directions, corresponding to 34, 37, 15, and 15 clear pairs per query. These full-cohort rows must not be subtracted from the smaller prompt-revision subset.

Model	Query	N	00	10	01	11
Phi-4-mini	Floral	68	3/2/63	19/5/44	23/18/27	30/25/13
	Roasted	74	6/5/63	7/6/61	29/20/25	34/19/21
	Caramellic	30	0/2/28	3/2/25	3/3/24	7/6/17
	Meaty	30	1/0/29	4/3/23	12/7/11	16/8/6
Qwen3-4B	Floral	68	33/35/0	33/35/0	37/31/0	34/34/0
	Roasted	74	44/28/2	36/36/2	43/29/2	36/36/2
	Caramellic	30	17/11/2	14/13/3	16/12/2	14/14/2
	Meaty	30	17/11/2	15/13/2	17/11/2	15/13/2

establish whether the factor effects persist under that wording regime. A conclusion about these original-prompt cells is not a general absence-of-benefit claim about structure or dispersion.

Phi's dispersion effects on correctness are $+22.79$, $+33.78$, $+11.67$, and $+38.33$ percentage points for floral, roasted, caramellic, and meaty. Each is accompanied by increased explicit errors and coverage. Qwen3-4B's corresponding correctness effects are $+3.68$, -0.68, -1.67, and 0.00 points. Its structure-visibility effect on roasted is -10.14 points, with an unadjusted paired interval of $[-16.22, -4.73]$; this is an average factorial effect, not the difference between two arbitrarily selected cells.

Under the full-information condition, Phi's clear-meaty reference-relative assignment advantage against complete-block-swapped evidence is $+20.00$ points $[3.33, 33.33]$. It answers correctly in both orders on 3/15 pairs. Its accepted-answer risk is 8/24, compared with 4/24 for the raw-gap selector and 2/24 for the normalized-gap selector at matching coverage. This local evidence-response benefit is distinct from predictive or selective added value.

For these same full-cohort clear pairs, the visible-score rule yields 46/22/0 on floral, 58/14/2 on roasted, 16/12/2 on caramellic, and 24/4/2 on meaty, in correct/wrong/TIE order. These query-specific results establish the rule's empirical decision value relative to the fixed panel target; they do not establish an invariant sensory ground truth. More correct answers need not indicate a better selective policy. These findings do not rule out gains from structure or calibrated uncertainty in other interfaces. Donor-linked floral and roasted each have one component; their intervals are therefore unavailable.

F Prompt Text for the Matched Revision Experiment

This appendix documents the prompt revision evaluated in Section 5.6 and Figure 3. The passages below reproduce the system and context text from the archived experimental inputs. Molecular structures, evidence values, question wording, and output contract remained unchanged in each matched old/new input. These excerpts describe one joint revision; no result isolates a sentence-level effect. Complete messages and their hashes are retained in the archived input differences. The original system already distinguished prediction from reference and already warned that ensemble SD was

not calibrated uncertainty. The revision changes the location and salience of those qualifications, removes the explicit transfer warning, adds condition-specific evidence labels, and emphasizes the output instruction. Its observed effect cannot be assigned separately to uncertainty interpretation, reduced caution, or improved instruction following.

Original system, shared by both conditions.

> You compare two molecules using the visible information. A model mean is a predicted descriptor association, not a calibrated probability or a human rating. Ensemble SD is dispersion across five fitted models, not an error probability, confidence interval, or guarantee. Use only the available information. You may decline to select a molecule by answering TIE. When reference measurements are provided, they are a label-assisted positive control, not model predictions.

Revised system, shared by both conditions.

> You compare two molecules using the visible information. Use only the available information. You may decline to select a molecule by answering TIE. Follow the output contract.

Shared context, unchanged.

> Task: compare published human panel mean descriptor applicability ratings (0-5) under each sample's reported preparation conditions. Zero means not selected in a rating, not physical absence.

Original context continuation, used even in the reference condition.

> Predictions below were trained on database descriptor annotations, so transfer to these ratings is uncertain.

Revised prediction-condition continuation.

> Evidence type: model predictions. model_mean is the mean of five fixed predictors trained on database descriptor annotations; it is not a human panel rating or a calibrated probability. ensemble_sd is the population standard deviation across those five predictors; it is not a calibrated error probability, a confidence interval, or a guarantee.

Revised reference-condition continuation.

> Evidence type: reference measurements. The reference_panel_mean_applicability values below are the published human panel means for this query and these samples, on the 0-5 rating scale. They are observed reference ratings, not model predictions. This is a label-assisted positive control.

Output contract, unchanged.

> Output contract: On the first line, output exactly A, B, or TIE. Do not output an explanation. TIE means that you decline to select A or B; it does not assert that the true ratings or associations are equal.

G GENERATION ACCOUNTING

Table 9 gives the stage totals. The original five stages contributed 24,426 formal generation events; the label/order follow-up added 3,232, giving 27,658 in total. A generation event is an execution-cost unit, not an independent molecule, pair, or distinct prompt. Repeated inputs and reruns remain in this event count. Historical answers used again in analysis do not enter it a second time.

Follow-up accounting. The label/order stage made 3,312 new attempts: 3,232 formal generations, 32 technical diagnostics, and 48 repeatability anchors. It reused 808 historical formal outputs from FF/FO without recounting their generation. There were no failed attempts or retries. The diagnostics

Table 9: Generation inventory. Formal events, technical diagnostics, and repeatability anchors are separate counts. Models, encodings, and repeated observations do not increase the number of independent molecular tasks. The factorial stage uses full cohorts for Phi/4B and fixed subsets for 8B.

Stage	New formal events	Diagnostics	Repeated anchors
Original	6,336	0	0
Extension	5,460	80	0
Factorial	11,350	48	0
Prompt comparison	960	0	0
32B reference	320	16	0
Label/order controls	3,232	32	48
Total	27,658	176	48

checked the numerical/output interface and were not used to select a configuration by accuracy; the anchors checked repeatability. All 48 matched their archived raw text and tokens. The LF historical runtime-lock hash could not be reconciled with the retained runtime file, so its historical-reuse option was rejected and its formal outputs were regenerated. Matching anchors alone do not establish that every historical runtime was reconstructed.

Historical LF versus regenerated E1/E4. We compare all 202 shared presentations per model by pair identity, direction, complete messages, and the label-to-molecule mapping, then parse both raw outputs independently. Physical choices, raw response text, and generated token sequences agree on 202/202 presentations for both Qwen3-4B and Phi. Neither model has a changed pair (0/101). The sets of rule-correct pairs lost in at least one original order are also identical: 11/11 for Qwen and 26/26 for Phi (Table 10). Thus the historical LF decision ledger corresponds directly to the regenerated four-encoding baseline at the output level. The four-encoding analysis still uses regenerated LF records; this equality does not resolve the historical runtime-lock discrepancy.

Table 10: Historical LF versus regenerated E1/E4. Physical choices, raw text, and generated tokens agree for every shared presentation. A changed pair has at least one changed physical decision. Equality of observed outputs does not resolve the historical runtime-lock discrepancy.

Model	Same physical choice	Changed pairs	Shared lost rule-correct pairs
Phi	202/202	0/101	26/26
Qwen3-4B	202/202	0/101	11/11

The formal follow-up contains 4,040 call identifiers (3,232 new and 808 reused), which map to 3,996 distinct model-plus-ordered-message input keys. The latter key excludes runtime configuration; configuration agreement is documented separately by fingerprints and reuse decisions. The difference of 44 comes entirely from repeated numerical inputs: 40 two-ID groups have identical displayed numbers across different pairs, and four two-ID groups arise when equal values make two encodings identical. These executed calls remain generation events, but are not distinct scientific questions.

There are 4,848 formal analysis uses: 2,424 scientific and 2,424 numerical. LF and FF use the same 808 numerical outputs, so this adds analysis uses without calls or independent replication. The frozen plan's field named `unique_formal_inputs_model_specific` records 4,040 planned call identifiers; strict message deduplication yields 3,996. The archived field is retained with this clarification. None of these accounting totals is used as the effective independent sample size.

Original five-stage detail. The following detailed ledger is preserved for the original five stages only. Its 24,426-new-response total is the historical subtotal in Table 9; it excludes the subsequent label/order controls described above. Rows expand model, query, condition, presentation order, and reuse status. Historical answers appearing as comparison material are not charged again.

Table 11: Formal-response generation ledger. Each row represents every listed model, condition, and both presentation orders. n is the number of responses per model, condition, and order; N is the resulting row total. Reused answers retain their original generation stage and are excluded from the 24,426-new-response total.

Stage	Source	Models	Query	Conditions	Role	n	N
Initial	Dev.	Q1.7,Phi,Q4	bergamot	$\mathcal{I}$	new	2	72
Initial	Dev.	Q1.7,Phi,Q4	brandy	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	cabbage	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	caramellic	$\mathcal{I}$	new	2	72
Initial	Dev.	Q1.7,Phi,Q4	cedar	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	chamomile	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	cherry	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	coconut	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	cortex	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	ethereal	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	floral	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	gassy	$\mathcal{I}$	new	2	72
Initial	Dev.	Q1.7,Phi,Q4	grape	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	hawthorn	$\mathcal{I}$	new	2	72
Initial	Dev.	Q1.7,Phi,Q4	meaty	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	natural	$\mathcal{I}$	new	2	72
Initial	Dev.	Q1.7,Phi,Q4	onion	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	orange	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	plum	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	popcorn	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	roasted	$\mathcal{I}$	new	1	36
Initial	Dev.	Q1.7,Phi,Q4	savory	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	smoky	$\mathcal{I}$	new	2	72
Initial	Dev.	Q1.7,Phi,Q4	sour	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	terpenic	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	tropical	$\mathcal{I}$	new	3	108
Initial	Dev.	Q1.7,Phi,Q4	violet	$\mathcal{I}$	new	2	72
Initial	Dev.	Q1.7,Phi,Q4	warm	$\mathcal{I}$	new	3	108
Initial	Ext.	Q1.7,Phi,Q4	floral	$\mathcal{I}$	new	101	3636
Extension	Ext.	Q8	floral	$\mathcal{I}$	new	101	1212
Extension	Ext.	Phi,Q4,Q8	nutty	$\mathcal{I}$	new	17	612
Extension	Ext.	Phi,Q4,Q8	odorless	$\mathcal{I}$	new	101	3636
Extension	Ext.	Q1.7,Phi,Q4	floral	$\mathcal{I}$	reused	101	3636
Factorial	Released	Phi,Q4	floral	$\mathcal{F}$	new	101	2020
Factorial	Released	Q8	floral	$\mathcal{F}$	new	20	200
Factorial	Released	Phi,Q4,Q8	nutty	$\mathcal{F}$	new	17	510
Factorial	Released	Phi,Q4	odorless	$\mathcal{F}$	new	101	2020
Factorial	Released	Q8	odorless	$\mathcal{F}$	new	20	200
Factorial	Human	Phi,Q4	caramellic	$\mathcal{H}$	new	40	1280
Factorial	Human	Q8	caramellic	$\mathcal{H}$	new	20	320
Factorial	Human	Phi,Q4	floral	$\mathcal{H}$	new	40	1280
Factorial	Human	Q8	floral	$\mathcal{H}$	new	20	320
Factorial	Human	Phi,Q4	meaty	$\mathcal{H}$	new	40	1280
Factorial	Human	Q8	meaty	$\mathcal{H}$	new	20	320
Factorial	Human	Phi,Q4	roasted	$\mathcal{H}$	new	40	1280
Factorial	Human	Q8	roasted	$\mathcal{H}$	new	20	320
Revision	Human	Phi,Q4,Q8	caramellic	$\mathcal{R}$	new	20	240
Revision	Human	Phi,Q4,Q8	floral	$\mathcal{R}$	new	20	240
Revision	Human	Phi,Q4,Q8	meaty	$\mathcal{R}$	new	20	240
Revision	Human	Phi,Q4,Q8	roasted	$\mathcal{R}$	new	20	240
Revision	Human	Phi,Q4,Q8	caramellic	$\mathcal{R}$	reused	20	240
Revision	Human	Phi,Q4,Q8	floral	$\mathcal{R}$	reused	20	240
Revision	Human	Phi,Q4,Q8	meaty	$\mathcal{R}$	reused	20	240
Revision	Human	Phi,Q4,Q8	roasted	$\mathcal{R}$	reused	20	240
32B	Human	Q32	caramellic	$\mathcal{R}$	new	20	80
32B	Human	Q32	floral	$\mathcal{R}$	new	20	80
32B	Human	Q32	meaty	$\mathcal{R}$	new	20	80
32B	Human	Q32	roasted	$\mathcal{R}$	new	20	80

Continued on next page.

Table 11 continued.

Stage	Source	Models	Query	Conditions	Role	n	N
32B	Human	Phi,Q4,Q8	caramellic	$\mathcal{R}$	reused	20	240
32B	Human	Phi,Q4,Q8	floral	$\mathcal{R}$	reused	20	240
32B	Human	Phi,Q4,Q8	meaty	$\mathcal{R}$	reused	20	240
32B	Human	Phi,Q4,Q8	roasted	$\mathcal{R}$	reused	20	240

Q1.7, Q4, Q8, and Q32 denote the archived Qwen configurations; Phi denotes Phi-4-mini.

$\mathcal{I} = \{\mathrm{N}, \mathrm{I}, \text{I-swap}, \mathrm{D}, \text{D-swap}, \mathrm{O}\}$; $\mathcal{F} = \{00, 10, 01, 11, \mathrm{SDperm}\}$; $\mathcal{H} = \mathcal{F} \cup \{\mathrm{N}, \mathrm{O}, \mathrm{SWAP}\}$; $\mathcal{R} = \{\mathrm{O}, 11\}$. Factorial digits denote SMILES identity and ensemble-SD visibility, respectively. Extension reuse comes from Initial, Revision reuse from Factorial, and 32B-stage reuse from Revision. All listed cells contain successful saved executions; legal TIE and invalid answers remain in the formal-response counts. Diagnostics are excluded.

H Displayed-Score Audit and Initial Decision Retention

The direct and swapped rule results in Table 1 were recomputed from the exact three-decimal `model_score` strings in the archived rendered inputs, using decimal comparison. No hidden higher-precision predictor output enters this rule. The 101 initial floral pairs contain no displayed ties: direct evidence yields 156 correct and 46 wrong presentations, while swapped evidence yields 46 correct and 156 wrong. The reference-relative assignment advantage is therefore $110/202 = 54.46$ percentage points. This is an empirical point estimate on the fixed pairs; no new interval is attached to it. The rule's net increment over itself is zero by definition.

Table 12: Initial floral retention and net correctness under direct evidence. Retention denominators are the 156 rule-correct presentations and 78 rule-correct pairs. Stable correction uses the 23 explicitly rule-wrong pairs; there are no rule ties. G and L are presentation counts. All models answer every presentation in this comparison, so every lost correct choice becomes an explicit wrong choice.

Model	Retained	Stable retained	Stable correction	G	L	$G - L$
Qwen3-1.7B	78/156	0/78	0/23	23	78	-55
Phi-4-mini	130/156	52/78	0/23	12	26	-14
Qwen3-4B	145/156	67/78	0/23	13	11	$+2$

For Qwen3-4B, the four pair groups are given in Table 2. Phi retains both correct decisions on 52 rule-correct pairs and loses the forward decision on the other 26. Of its 23 rule-wrong pairs, 12 become correct only in the reversed presentation and 11 remain wrong in both. Its 38 unstable pairs choose position B twice. Qwen3-1.7B chooses A on every presentation. Thus none of these initial models stably corrects a rule-wrong pair in both orders, although all have favorable presentation-level transitions. This does not establish a unique mechanism: the outcome patterns are compatible with position-dependent selection.

The recomputation cross-checks 1,212 existing direct/swapped model records across the three configurations against saved raw responses and input/output hashes. The source records, extraction script, displayed-score ledger, per-presentation outcomes, and pair-level transition ledger accompany the revision. It creates no model responses and does not change historical confidence intervals.

H.1 Physical response to evidence reassignment

The decision-change rate in Equation 1 uses physical molecule identity, not correctness. We pair D and D-swap by molecule pair and presentation direction, verify unchanged identities and exchanged displayed means/dispersion, parse the first nonempty output line, and map A/B back to m_1/m_2. The four decision states are m_1, m_2, `TIE`, and invalid. A switch to or from abstention would count as a state change; no abstentions or invalid outputs occur in this comparison. The 202 matched presentations reuse 101 pairs in both orders.

Table 13: Reference-free response to direct-evidence reassignment, split by presentation direction. Numerators count physical decision-state changes, not favorable changes. The final column is a separate reference-relative correctness contrast. Historical confidence intervals in Table 1 are unchanged; no new resampling is performed.

Model	Forward	Reversed	Total changes	Δ_{assign} (pp)
Qwen3-1.7B	0/101	0/101	0/202	0.00
Phi-4-mini	69/101	50/101	119/202	31.19
Qwen3-4B	73/101	85/101	158/202	50.50
Visible-score rule	101/101	101/101	202/202	54.46

Every displayed-score choice reverses because the means never tie, whereas model responses change less often. An always-following binary rule with reference accuracy p has change rate one but assignment advantage $p-(1-p) = 2p-1$. At $p = 1/2$ the advantage is zero despite complete behavioral response. Thus assignment advantage, physical response, retained reference-correct decisions, and added correctness require separate reporting. The accompanying transition ledger retains all four states and all matched presentations; this analysis makes no new model calls.

I Complete archived outcomes and matched prompt transitions

This appendix reports all frozen conditions from the factorial study, all matched old/new outcomes on the fixed 80-pair human subset, and the 32B reference run. The portable audit reconstructs the matched human-interface and 32B counts from saved first-line responses. The full-cohort factorial tables are frozen aggregate exports; the included response replay covers only part of those cohorts and does not independently verify every factorial cell. C denotes correct, W an explicit wrong A/B choice, T abstention (`TIE`), and I an invalid answer. Human *clear* and *ambiguous* rows are disjoint; both use the frozen panel-mean ordering as the reference. All outcome cells retain the full presentation denominator. Two orders of a pair are repeated observations, not independent samples. These supplementary counts do not introduce additional hypotheses or confidence intervals. The electronic tables retain every presentation direction and the complete four-by-four transition matrices.

I.1 Complete factorial and control outcomes

Table 14: Complete released-label factorial and SD-permutation outcomes under the original factorial prompt. N is the number of presentations, not independent molecules. Phi-mini and Qwen-4B use all 101/101/17 pairs; Qwen-8B uses its fixed 20/20/17-pair subset (floral/odorless/nutty).

Model	Query	Scope	Cond.	N	C	W	T	I
Phi-mini	floral	released	00	202	105	97	0	0
Phi-mini	floral	released	10	202	148	54	0	0
Phi-mini	floral	released	01	202	145	57	0	0
Phi-mini	floral	released	11	202	152	50	0	0
Phi-mini	floral	released	SD-perm	202	154	48	0	0
Phi-mini	odorless	released	00	202	116	84	2	0
Phi-mini	odorless	released	10	202	112	88	2	0
Phi-mini	odorless	released	01	202	125	77	0	0
Phi-mini	odorless	released	11	202	101	99	2	0
Phi-mini	odorless	released	SD-perm	202	104	96	2	0
Phi-mini	nutty	released	00	34	17	17	0	0
Phi-mini	nutty	released	10	34	25	9	0	0
Phi-mini	nutty	released	01	34	24	10	0	0
Phi-mini	nutty	released	11	34	26	8	0	0
Phi-mini	nutty	released	SD-perm	34	27	7	0	0
Qwen-4B	floral	released	00	202	156	46	0	0
Qwen-4B	floral	released	10	202	156	46	0	0
Qwen-4B	floral	released	01	202	158	44	0	0

Continued on next page

Table 14 (continued)

Model	Query	Scope	Cond.	N	C	W	T	I
Qwen-4B	floral	released	11	202	158	44	0	0
Qwen-4B	floral	released	SD-perm	202	160	42	0	0
Qwen-4B	odorless	released	00	202	118	82	2	0
Qwen-4B	odorless	released	10	202	112	88	2	0
Qwen-4B	odorless	released	01	202	105	95	2	0
Qwen-4B	odorless	released	11	202	102	98	2	0
Qwen-4B	odorless	released	SD-perm	202	99	101	2	0
Qwen-4B	nutty	released	00	34	26	8	0	0
Qwen-4B	nutty	released	10	34	26	8	0	0
Qwen-4B	nutty	released	01	34	25	9	0	0
Qwen-4B	nutty	released	11	34	25	9	0	0
Qwen-4B	nutty	released	SD-perm	34	25	9	0	0
Qwen-8B	floral	released	00	40	28	11	1	0
Qwen-8B	floral	released	10	40	27	10	3	0
Qwen-8B	floral	released	01	40	28	11	1	0
Qwen-8B	floral	released	11	40	27	11	2	0
Qwen-8B	floral	released	SD-perm	40	28	12	0	0
Qwen-8B	odorless	released	00	40	30	8	2	0
Qwen-8B	odorless	released	10	40	17	7	16	0
Qwen-8B	odorless	released	01	40	22	18	0	0
Qwen-8B	odorless	released	11	40	20	12	8	0
Qwen-8B	odorless	released	SD-perm	40	23	12	5	0
Qwen-8B	nutty	released	00	34	25	8	1	0
Qwen-8B	nutty	released	10	34	24	8	2	0
Qwen-8B	nutty	released	01	34	25	8	1	0
Qwen-8B	nutty	released	11	34	25	7	2	0
Qwen-8B	nutty	released	SD-perm	34	26	8	0	0

Table 15: Complete human-reference factorial, SD-permutation, no-evidence (N), complete-block swap, and reference-reading (O) outcomes under the original prompt. Phi-mini and Qwen-4B use 40 pairs per query before clear/ambiguous separation; Qwen-8B uses the fixed 20-pair subset. O gives target reference means and is not prediction accuracy.

Model	Query	Scope	Cond.	N	C	W	T	I
Phi-mini	floral	clear	00	68	3	2	63	0
Phi-mini	floral	clear	10	68	19	5	44	0
Phi-mini	floral	clear	01	68	23	18	27	0
Phi-mini	floral	clear	11	68	30	25	13	0
Phi-mini	floral	clear	SD-perm	68	32	23	13	0
Phi-mini	floral	clear	N	68	0	0	68	0
Phi-mini	floral	clear	Swap	68	23	32	13	0
Phi-mini	floral	clear	O	68	68	0	0	0
Phi-mini	floral	ambiguous	00	12	2	1	9	0
Phi-mini	floral	ambiguous	10	12	2	3	7	0
Phi-mini	floral	ambiguous	01	12	4	2	6	0
Phi-mini	floral	ambiguous	11	12	6	3	3	0
Phi-mini	floral	ambiguous	SD-perm	12	7	2	3	0
Phi-mini	floral	ambiguous	N	12	0	0	12	0
Phi-mini	floral	ambiguous	Swap	12	4	5	3	0
Phi-mini	floral	ambiguous	O	12	10	2	0	0
Phi-mini	roasted	clear	00	74	6	5	63	0
Phi-mini	roasted	clear	10	74	7	6	61	0
Phi-mini	roasted	clear	01	74	29	20	25	0
Phi-mini	roasted	clear	11	74	34	19	21	0
Phi-mini	roasted	clear	SD-perm	74	34	13	27	0
Phi-mini	roasted	clear	N	74	0	0	74	0

Continued on next page

Table 15 (continued)

Model	Query	Scope	Cond.	N	C	W	T	I
Phi-mini	roasted	clear	Swap	74	27	26	21	0
Phi-mini	roasted	clear	O	74	72	2	0	0
Phi-mini	roasted	ambiguous	00	6	0	0	6	0
Phi-mini	roasted	ambiguous	10	6	0	0	6	0
Phi-mini	roasted	ambiguous	01	6	1	1	4	0
Phi-mini	roasted	ambiguous	11	6	1	1	4	0
Phi-mini	roasted	ambiguous	SD-perm	6	0	1	5	0
Phi-mini	roasted	ambiguous	N	6	0	0	6	0
Phi-mini	roasted	ambiguous	Swap	6	1	2	3	0
Phi-mini	roasted	ambiguous	O	6	5	1	0	0
Phi-mini	caramellic	clear	00	30	0	2	28	0
Phi-mini	caramellic	clear	10	30	3	2	25	0
Phi-mini	caramellic	clear	01	30	3	3	24	0
Phi-mini	caramellic	clear	11	30	7	6	17	0
Phi-mini	caramellic	clear	SD-perm	30	5	6	19	0
Phi-mini	caramellic	clear	N	30	0	0	30	0
Phi-mini	caramellic	clear	Swap	30	7	7	16	0
Phi-mini	caramellic	clear	O	30	30	0	0	0
Phi-mini	caramellic	ambiguous	00	50	3	3	44	0
Phi-mini	caramellic	ambiguous	10	50	5	2	43	0
Phi-mini	caramellic	ambiguous	01	50	14	5	31	0
Phi-mini	caramellic	ambiguous	11	50	15	15	20	0
Phi-mini	caramellic	ambiguous	SD-perm	50	15	12	23	0
Phi-mini	caramellic	ambiguous	N	50	0	0	50	0
Phi-mini	caramellic	ambiguous	Swap	50	14	15	21	0
Phi-mini	caramellic	ambiguous	O	50	43	7	0	0
Phi-mini	meaty	clear	00	30	1	0	29	0
Phi-mini	meaty	clear	10	30	4	3	23	0
Phi-mini	meaty	clear	01	30	12	7	11	0
Phi-mini	meaty	clear	11	30	16	8	6	0
Phi-mini	meaty	clear	SD-perm	30	15	9	6	0
Phi-mini	meaty	clear	N	30	0	0	30	0
Phi-mini	meaty	clear	Swap	30	10	15	5	0
Phi-mini	meaty	clear	O	30	29	1	0	0
Phi-mini	meaty	ambiguous	00	50	3	1	46	0
Phi-mini	meaty	ambiguous	10	50	9	5	36	0
Phi-mini	meaty	ambiguous	01	50	17	12	21	0
Phi-mini	meaty	ambiguous	11	50	18	19	13	0
Phi-mini	meaty	ambiguous	SD-perm	50	21	17	12	0
Phi-mini	meaty	ambiguous	N	50	0	0	50	0
Phi-mini	meaty	ambiguous	Swap	50	17	18	15	0
Phi-mini	meaty	ambiguous	O	50	47	3	0	0
Qwen-4B	floral	clear	00	68	33	35	0	0
Qwen-4B	floral	clear	10	68	33	35	0	0
Qwen-4B	floral	clear	01	68	37	31	0	0
Qwen-4B	floral	clear	11	68	34	34	0	0
Qwen-4B	floral	clear	SD-perm	68	36	32	0	0
Qwen-4B	floral	clear	N	68	0	0	68	0
Qwen-4B	floral	clear	Swap	68	33	35	0	0
Qwen-4B	floral	clear	O	68	68	0	0	0
Qwen-4B	floral	ambiguous	00	12	9	3	0	0
Qwen-4B	floral	ambiguous	10	12	7	5	0	0
Qwen-4B	floral	ambiguous	01	12	7	5	0	0
Qwen-4B	floral	ambiguous	11	12	5	7	0	0
Qwen-4B	floral	ambiguous	SD-perm	12	5	7	0	0
Qwen-4B	floral	ambiguous	N	12	0	0	12	0
Qwen-4B	floral	ambiguous	Swap	12	6	6	0	0
Qwen-4B	floral	ambiguous	O	12	11	1	0	0

Continued on next page

Table 15 (continued)

Model	Query	Scope	Cond.	N	C	W	T	I
Qwen-4B	roasted	clear	00	74	44	28	2	0
Qwen-4B	roasted	clear	10	74	36	36	2	0
Qwen-4B	roasted	clear	01	74	43	29	2	0
Qwen-4B	roasted	clear	11	74	36	36	2	0
Qwen-4B	roasted	clear	SD-perm	74	38	36	0	0
Qwen-4B	roasted	clear	N	74	0	0	74	0
Qwen-4B	roasted	clear	Swap	74	36	36	2	0
Qwen-4B	roasted	clear	O	74	71	3	0	0
Qwen-4B	roasted	ambiguous	00	6	3	3	0	0
Qwen-4B	roasted	ambiguous	10	6	3	3	0	0
Qwen-4B	roasted	ambiguous	01	6	3	3	0	0
Qwen-4B	roasted	ambiguous	11	6	3	3	0	0
Qwen-4B	roasted	ambiguous	SD-perm	6	3	3	0	0
Qwen-4B	roasted	ambiguous	N	6	0	0	6	0
Qwen-4B	roasted	ambiguous	Swap	6	3	3	0	0
Qwen-4B	roasted	ambiguous	O	6	6	0	0	0
Qwen-4B	caramellic	clear	00	30	17	11	2	0
Qwen-4B	caramellic	clear	10	30	14	13	3	0
Qwen-4B	caramellic	clear	01	30	16	12	2	0
Qwen-4B	caramellic	clear	11	30	14	14	2	0
Qwen-4B	caramellic	clear	SD-perm	30	14	14	2	0
Qwen-4B	caramellic	clear	N	30	0	0	30	0
Qwen-4B	caramellic	clear	Swap	30	14	14	2	0
Qwen-4B	caramellic	clear	O	30	30	0	0	0
Qwen-4B	caramellic	ambiguous	00	50	23	23	4	0
Qwen-4B	caramellic	ambiguous	10	50	21	24	5	0
Qwen-4B	caramellic	ambiguous	01	50	20	26	4	0
Qwen-4B	caramellic	ambiguous	11	50	22	24	4	0
Qwen-4B	caramellic	ambiguous	SD-perm	50	22	24	4	0
Qwen-4B	caramellic	ambiguous	N	50	0	0	50	0
Qwen-4B	caramellic	ambiguous	Swap	50	24	22	4	0
Qwen-4B	caramellic	ambiguous	O	50	45	5	0	0
Qwen-4B	meaty	clear	00	30	17	11	2	0
Qwen-4B	meaty	clear	10	30	15	13	2	0
Qwen-4B	meaty	clear	01	30	17	11	2	0
Qwen-4B	meaty	clear	11	30	15	13	2	0
Qwen-4B	meaty	clear	SD-perm	30	15	13	2	0
Qwen-4B	meaty	clear	N	30	0	0	30	0
Qwen-4B	meaty	clear	Swap	30	14	14	2	0
Qwen-4B	meaty	clear	O	30	22	8	0	0
Qwen-4B	meaty	ambiguous	00	50	26	22	2	0
Qwen-4B	meaty	ambiguous	10	50	25	22	3	0
Qwen-4B	meaty	ambiguous	01	50	26	22	2	0
Qwen-4B	meaty	ambiguous	11	50	25	23	2	0
Qwen-4B	meaty	ambiguous	SD-perm	50	24	24	2	0
Qwen-4B	meaty	ambiguous	N	50	0	0	50	0
Qwen-4B	meaty	ambiguous	Swap	50	24	24	2	0
Qwen-4B	meaty	ambiguous	O	50	26	24	0	0
Qwen-8B	floral	clear	00	36	0	0	36	0
Qwen-8B	floral	clear	10	36	0	0	36	0
Qwen-8B	floral	clear	01	36	0	0	36	0
Qwen-8B	floral	clear	11	36	0	0	36	0
Qwen-8B	floral	clear	SD-perm	36	0	0	36	0
Qwen-8B	floral	clear	N	36	0	0	36	0
Qwen-8B	floral	clear	Swap	36	0	0	36	0
Qwen-8B	floral	clear	O	36	23	0	13	0
Qwen-8B	floral	ambiguous	00	4	0	0	4	0
Qwen-8B	floral	ambiguous	10	4	0	0	4	0

Continued on next page

Table 15 (continued)

Model	Query	Scope	Cond.	N	C	W	T	I
Qwen-8B	floral	ambiguous	01	4	0	0	4	0
Qwen-8B	floral	ambiguous	11	4	0	0	4	0
Qwen-8B	floral	ambiguous	SD-perm	4	0	0	4	0
Qwen-8B	floral	ambiguous	N	4	0	0	4	0
Qwen-8B	floral	ambiguous	Swap	4	0	0	4	0
Qwen-8B	floral	ambiguous	O	4	2	0	2	0
Qwen-8B	roasted	clear	00	38	3	0	35	0
Qwen-8B	roasted	clear	10	38	0	0	38	0
Qwen-8B	roasted	clear	01	38	2	0	36	0
Qwen-8B	roasted	clear	11	38	0	0	38	0
Qwen-8B	roasted	clear	SD-perm	38	0	0	38	0
Qwen-8B	roasted	clear	N	38	0	0	38	0
Qwen-8B	roasted	clear	Swap	38	0	0	38	0
Qwen-8B	roasted	clear	O	38	0	0	38	0
Qwen-8B	roasted	ambiguous	00	2	0	0	2	0
Qwen-8B	roasted	ambiguous	10	2	0	0	2	0
Qwen-8B	roasted	ambiguous	01	2	0	0	2	0
Qwen-8B	roasted	ambiguous	11	2	0	0	2	0
Qwen-8B	roasted	ambiguous	SD-perm	2	0	0	2	0
Qwen-8B	roasted	ambiguous	N	2	0	0	2	0
Qwen-8B	roasted	ambiguous	Swap	2	0	0	2	0
Qwen-8B	roasted	ambiguous	O	2	0	0	2	0
Qwen-8B	caramellic	clear	00	14	0	0	14	0
Qwen-8B	caramellic	clear	10	14	0	0	14	0
Qwen-8B	caramellic	clear	01	14	0	0	14	0
Qwen-8B	caramellic	clear	11	14	0	0	14	0
Qwen-8B	caramellic	clear	SD-perm	14	0	0	14	0
Qwen-8B	caramellic	clear	N	14	0	0	14	0
Qwen-8B	caramellic	clear	Swap	14	0	0	14	0
Qwen-8B	caramellic	clear	O	14	4	0	10	0
Qwen-8B	caramellic	ambiguous	00	26	0	0	26	0
Qwen-8B	caramellic	ambiguous	10	26	0	0	26	0
Qwen-8B	caramellic	ambiguous	01	26	0	0	26	0
Qwen-8B	caramellic	ambiguous	11	26	0	0	26	0
Qwen-8B	caramellic	ambiguous	SD-perm	26	0	0	26	0
Qwen-8B	caramellic	ambiguous	N	26	0	0	26	0
Qwen-8B	caramellic	ambiguous	Swap	26	0	0	26	0
Qwen-8B	caramellic	ambiguous	O	26	0	0	26	0
Qwen-8B	meaty	clear	00	14	0	0	14	0
Qwen-8B	meaty	clear	10	14	0	0	14	0
Qwen-8B	meaty	clear	01	14	0	0	14	0
Qwen-8B	meaty	clear	11	14	0	0	14	0
Qwen-8B	meaty	clear	SD-perm	14	0	0	14	0
Qwen-8B	meaty	clear	N	14	0	0	14	0
Qwen-8B	meaty	clear	Swap	14	0	0	14	0
Qwen-8B	meaty	clear	O	14	2	0	12	0
Qwen-8B	meaty	ambiguous	00	26	0	0	26	0
Qwen-8B	meaty	ambiguous	10	26	0	0	26	0
Qwen-8B	meaty	ambiguous	01	26	0	0	26	0
Qwen-8B	meaty	ambiguous	11	26	0	0	26	0
Qwen-8B	meaty	ambiguous	SD-perm	26	0	0	26	0
Qwen-8B	meaty	ambiguous	N	26	0	0	26	0
Qwen-8B	meaty	ambiguous	Swap	26	0	0	26	0
Qwen-8B	meaty	ambiguous	O	26	0	0	26	0

I.2 Reference-relative score performance

Table 16 evaluates the actual displayed three-decimal argmax against each frozen human-reference subset. It describes score information in this selected high/low-panel-mean diagnostic distribution.

Ranking avoids a numerical-scale comparison but does not equate database associations with applicability ratings under different preparations. The fixed subset is reused in the prompt revision and 32B run; it is not an additional independent sample.

Table 16: Displayed-score argmax outcomes against human panel-mean order. C/W/T/I count both presentations of each pair; numerical ties remain abstentions.

Sample	Query	Scope	Pairs	C	W	T	I
fixed20	floral	ambiguous	2	4	0	0	0
fixed20	floral	clear	18	28	8	0	0
fixed20	roasted	ambiguous	1	0	2	0	0
fixed20	roasted	clear	19	32	6	0	0
fixed20	caramellic	ambiguous	13	16	8	2	0
fixed20	caramellic	clear	7	6	6	2	0
fixed20	meaty	ambiguous	13	16	8	2	0
fixed20	meaty	clear	7	12	0	2	0
full40	floral	ambiguous	6	10	2	0	0
full40	floral	clear	34	46	22	0	0
full40	roasted	ambiguous	3	4	2	0	0
full40	roasted	clear	37	58	14	2	0
full40	caramellic	ambiguous	25	28	18	4	0
full40	caramellic	clear	15	16	12	2	0
full40	meaty	ambiguous	25	36	12	2	0
full40	meaty	clear	15	24	4	2	0

I.3 Matched prompt-revision transitions

Table 17 aligns each saved original/revised answer by model, pair, condition, and presentation direction. For example, CW is correct under the original wording and explicitly wrong under the revised wording; TC is original abstention followed by a correct choice. Thus a net increase in correct responses can contain gains, losses, and newly answered cases. All transitions involving I were zero; the accompanying CSV also reports those seven cells explicitly. The two orders are shown together here for compactness and separately in the CSV. The intervention was one joint wording/context package; these transitions do not isolate any sentence-level or psychological mechanism.

Table 17: Matched original-to-revised transitions: 11, clear. Every invalid-related cell is zero; N is the total presentations in this matched subset.

Model	Query	N	CC	CW	CT	WC	WW	WT	TC	TW	TT
Phi-mini	floral	36	15	3	0	9	4	0	4	1	0
Phi-mini	roasted	38	16	2	0	8	4	0	5	3	0
Phi-mini	caramellic	14	2	1	0	0	3	0	4	2	2
Phi-mini	meaty	14	6	0	0	5	0	0	1	0	2
Qwen-4B	floral	36	15	3	0	13	5	0	0	0	0
Qwen-4B	roasted	38	19	0	0	4	15	0	0	0	0
Qwen-4B	caramellic	14	5	1	0	2	4	0	0	0	2
Qwen-4B	meaty	14	6	0	0	3	3	0	0	0	2
Qwen-8B	floral	36	0	0	0	0	0	0	0	0	36
Qwen-8B	roasted	38	0	0	0	0	0	0	0	0	38
Qwen-8B	caramellic	14	0	0	0	0	0	0	0	0	14
Qwen-8B	meaty	14	0	0	0	0	0	0	0	0	14

Table 18: Matched original-to-revised transitions: 11, ambiguous. Every invalid-related cell is zero; N is the total presentations in this matched subset.

Model	Query	N	CC	CW	CT	WC	WW	WT	TC	TW	TT
Phi-mini	floral	4	2	0	0	1	0	0	1	0	0
Phi-mini	roasted	2	0	0	0	0	1	0	0	1	0
Phi-mini	caramellic	26	7	1	0	6	4	0	2	4	2
Phi-mini	meaty	26	7	1	0	6	4	0	4	2	2
Qwen-4B	floral	4	2	0	0	2	0	0	0	0	0
Qwen-4B	roasted	2	1	0	0	0	1	0	0	0	0
Qwen-4B	caramellic	26	10	2	0	1	11	0	0	0	2
Qwen-4B	meaty	26	12	0	0	2	10	0	0	1	1
Qwen-8B	floral	4	0	0	0	0	0	0	0	0	4
Qwen-8B	roasted	2	0	0	0	0	0	0	0	0	2
Qwen-8B	caramellic	26	0	0	0	0	0	0	0	0	26
Qwen-8B	meaty	26	0	0	0	0	0	0	0	0	26

Table 19: Matched original-to-revised transitions: O, clear. Every invalid-related cell is zero; N is the total presentations in this matched subset.

Model	Query	N	CC	CW	CT	WC	WW	WT	TC	TW	TT
Phi-mini	floral	36	36	0	0	0	0	0	0	0	0
Phi-mini	roasted	38	37	0	0	1	0	0	0	0	0
Phi-mini	caramellic	14	14	0	0	0	0	0	0	0	0
Phi-mini	meaty	14	13	0	0	0	1	0	0	0	0
Qwen-4B	floral	36	36	0	0	0	0	0	0	0	0
Qwen-4B	roasted	38	36	0	0	2	0	0	0	0	0
Qwen-4B	caramellic	14	14	0	0	0	0	0	0	0	0
Qwen-4B	meaty	14	9	0	0	3	2	0	0	0	0
Qwen-8B	floral	36	15	4	4	0	0	0	2	1	10
Qwen-8B	roasted	38	0	0	0	0	0	0	11	1	26
Qwen-8B	caramellic	14	4	0	0	0	0	0	1	0	9
Qwen-8B	meaty	14	2	0	0	0	0	0	5	0	7

Table 20: Matched original-to-revised transitions: O, ambiguous. Every invalid-related cell is zero; N is the total presentations in this matched subset.

Model	Query	N	CC	CW	CT	WC	WW	WT	TC	TW	TT
Phi-mini	floral	4	3	0	0	1	0	0	0	0	0
Phi-mini	roasted	2	1	0	0	1	0	0	0	0	0
Phi-mini	caramellic	26	23	0	0	3	0	0	0	0	0
Phi-mini	meaty	26	25	0	0	1	0	0	0	0	0
Qwen-4B	floral	4	4	0	0	0	0	0	0	0	0
Qwen-4B	roasted	2	2	0	0	0	0	0	0	0	0
Qwen-4B	caramellic	26	23	0	0	3	0	0	0	0	0
Qwen-4B	meaty	26	14	0	0	4	8	0	0	0	0
Qwen-8B	floral	4	1	1	0	0	0	0	0	0	2
Qwen-8B	roasted	2	0	0	0	0	0	0	0	0	2
Qwen-8B	caramellic	26	0	0	0	0	0	0	6	0	20
Qwen-8B	meaty	26	0	0	0	0	0	0	0	0	26

I.4 Complete fixed-subset outcomes

The following table includes all 29 ambiguous pairs omitted from the primary clear analysis (floral 2, roasted 1, caramellic 13, meaty 13), for every matched historical model and both conditions, as well as the 32B run. Clear rows are included to make the partition and full denominators visible. Original and revised refer to the prompt package; the 32B configuration has no original-prompt run. Repeated historical rows used in the 32B comparison are shown only once here. Ambiguous

reference-reading success indicates compliance with frozen mean order, not certainty about a new panel.

Table 21: All fixed-subset outcomes for the original/revised local configurations and revised 32B configuration. N is the presentation count for the displayed row, not an additional generation count.

Model	Query	Scope	Prompt	Cond.	N	C	W	T	I
Phi-mini	floral	clear	orig.	11	36	18	13	5	0
Phi-mini	floral	clear	rev.	11	36	28	8	0	0
Phi-mini	floral	clear	orig.	O	36	36	0	0	0
Phi-mini	floral	clear	rev.	O	36	36	0	0	0
Phi-mini	floral	ambiguous	orig.	11	4	2	1	1	0
Phi-mini	floral	ambiguous	rev.	11	4	4	0	0	0
Phi-mini	floral	ambiguous	orig.	O	4	3	1	0	0
Phi-mini	floral	ambiguous	rev.	O	4	4	0	0	0
Phi-mini	roasted	clear	orig.	11	38	18	12	8	0
Phi-mini	roasted	clear	rev.	11	38	29	9	0	0
Phi-mini	roasted	clear	orig.	O	38	37	1	0	0
Phi-mini	roasted	clear	rev.	O	38	38	0	0	0
Phi-mini	roasted	ambiguous	orig.	11	2	0	1	1	0
Phi-mini	roasted	ambiguous	rev.	11	2	0	2	0	0
Phi-mini	roasted	ambiguous	orig.	O	2	1	1	0	0
Phi-mini	roasted	ambiguous	rev.	O	2	2	0	0	0
Phi-mini	caramellic	clear	orig.	11	14	3	3	8	0
Phi-mini	caramellic	clear	rev.	11	14	6	6	2	0
Phi-mini	caramellic	clear	orig.	O	14	14	0	0	0
Phi-mini	caramellic	clear	rev.	O	14	14	0	0	0
Phi-mini	caramellic	ambiguous	orig.	11	26	8	10	8	0
Phi-mini	caramellic	ambiguous	rev.	11	26	15	9	2	0
Phi-mini	caramellic	ambiguous	orig.	O	26	23	3	0	0
Phi-mini	caramellic	ambiguous	rev.	O	26	26	0	0	0
Phi-mini	meaty	clear	orig.	11	14	6	5	3	0
Phi-mini	meaty	clear	rev.	11	14	12	0	2	0
Phi-mini	meaty	clear	orig.	O	14	13	1	0	0
Phi-mini	meaty	clear	rev.	O	14	13	1	0	0
Phi-mini	meaty	ambiguous	orig.	11	26	8	10	8	0
Phi-mini	meaty	ambiguous	rev.	11	26	17	7	2	0
Phi-mini	meaty	ambiguous	orig.	O	26	25	1	0	0
Phi-mini	meaty	ambiguous	rev.	O	26	26	0	0	0
Qwen-4B	floral	clear	orig.	11	36	18	18	0	0
Qwen-4B	floral	clear	rev.	11	36	28	8	0	0
Qwen-4B	floral	clear	orig.	O	36	36	0	0	0
Qwen-4B	floral	clear	rev.	O	36	36	0	0	0
Qwen-4B	floral	ambiguous	orig.	11	4	2	2	0	0
Qwen-4B	floral	ambiguous	rev.	11	4	4	0	0	0
Qwen-4B	floral	ambiguous	orig.	O	4	4	0	0	0
Qwen-4B	floral	ambiguous	rev.	O	4	4	0	0	0
Qwen-4B	roasted	clear	orig.	11	38	19	19	0	0
Qwen-4B	roasted	clear	rev.	11	38	23	15	0	0
Qwen-4B	roasted	clear	orig.	O	38	36	2	0	0
Qwen-4B	roasted	clear	rev.	O	38	38	0	0	0
Qwen-4B	roasted	ambiguous	orig.	11	2	1	1	0	0
Qwen-4B	roasted	ambiguous	rev.	11	2	1	1	0	0
Qwen-4B	roasted	ambiguous	orig.	O	2	2	0	0	0
Qwen-4B	roasted	ambiguous	rev.	O	2	2	0	0	0
Qwen-4B	caramellic	clear	orig.	11	14	6	6	2	0
Qwen-4B	caramellic	clear	rev.	11	14	7	5	2	0
Qwen-4B	caramellic	clear	orig.	O	14	14	0	0	0
Qwen-4B	caramellic	clear	rev.	O	14	14	0	0	0
Qwen-4B	caramellic	ambiguous	orig.	11	26	12	12	2	0

Continued on next page

Table 21 (continued)

Model	Query	Scope	Prompt	Cond.	N	C	W	T	I
Qwen-4B	caramellic	ambiguous	rev.	11	26	11	13	2	0
Qwen-4B	caramellic	ambiguous	orig.	O	26	23	3	0	0
Qwen-4B	caramellic	ambiguous	rev.	O	26	26	0	0	0
Qwen-4B	meaty	clear	orig.	11	14	6	6	2	0
Qwen-4B	meaty	clear	rev.	11	14	9	3	2	0
Qwen-4B	meaty	clear	orig.	O	14	9	5	0	0
Qwen-4B	meaty	clear	rev.	O	14	12	2	0	0
Qwen-4B	meaty	ambiguous	orig.	11	26	12	12	2	0
Qwen-4B	meaty	ambiguous	rev.	11	26	14	11	1	0
Qwen-4B	meaty	ambiguous	orig.	O	26	14	12	0	0
Qwen-4B	meaty	ambiguous	rev.	O	26	18	8	0	0
Qwen-8B	floral	clear	orig.	11	36	0	0	36	0
Qwen-8B	floral	clear	rev.	11	36	0	0	36	0
Qwen-8B	floral	clear	orig.	O	36	23	0	13	0
Qwen-8B	floral	clear	rev.	O	36	17	5	14	0
Qwen-8B	floral	ambiguous	orig.	11	4	0	0	4	0
Qwen-8B	floral	ambiguous	rev.	11	4	0	0	4	0
Qwen-8B	floral	ambiguous	orig.	O	4	2	0	2	0
Qwen-8B	floral	ambiguous	rev.	O	4	1	1	2	0
Qwen-8B	roasted	clear	orig.	11	38	0	0	38	0
Qwen-8B	roasted	clear	rev.	11	38	0	0	38	0
Qwen-8B	roasted	clear	orig.	O	38	0	0	38	0
Qwen-8B	roasted	clear	rev.	O	38	11	1	26	0
Qwen-8B	roasted	ambiguous	orig.	11	2	0	0	2	0
Qwen-8B	roasted	ambiguous	rev.	11	2	0	0	2	0
Qwen-8B	roasted	ambiguous	orig.	O	2	0	0	2	0
Qwen-8B	roasted	ambiguous	rev.	O	2	0	0	2	0
Qwen-8B	caramellic	clear	orig.	11	14	0	0	14	0
Qwen-8B	caramellic	clear	rev.	11	14	0	0	14	0
Qwen-8B	caramellic	clear	orig.	O	14	4	0	10	0
Qwen-8B	caramellic	clear	rev.	O	14	5	0	9	0
Qwen-8B	caramellic	ambiguous	orig.	11	26	0	0	26	0
Qwen-8B	caramellic	ambiguous	rev.	11	26	0	0	26	0
Qwen-8B	caramellic	ambiguous	orig.	O	26	0	0	26	0
Qwen-8B	caramellic	ambiguous	rev.	O	26	6	0	20	0
Qwen-8B	meaty	clear	orig.	11	14	0	0	14	0
Qwen-8B	meaty	clear	rev.	11	14	0	0	14	0
Qwen-8B	meaty	clear	orig.	O	14	2	0	12	0
Qwen-8B	meaty	clear	rev.	O	14	7	0	7	0
Qwen-8B	meaty	ambiguous	orig.	11	26	0	0	26	0
Qwen-8B	meaty	ambiguous	rev.	11	26	0	0	26	0
Qwen-8B	meaty	ambiguous	orig.	O	26	0	0	26	0
Qwen-8B	meaty	ambiguous	rev.	O	26	0	0	26	0
Qwen-32B	floral	clear	rev.	11	36	18	4	14	0
Qwen-32B	floral	clear	rev.	O	36	36	0	0	0
Qwen-32B	floral	ambiguous	rev.	11	4	1	0	3	0
Qwen-32B	floral	ambiguous	rev.	O	4	4	0	0	0
Qwen-32B	roasted	clear	rev.	11	38	5	0	33	0
Qwen-32B	roasted	clear	rev.	O	38	38	0	0	0
Qwen-32B	roasted	ambiguous	rev.	11	2	0	0	2	0
Qwen-32B	roasted	ambiguous	rev.	O	2	2	0	0	0
Qwen-32B	caramellic	clear	rev.	11	14	0	0	14	0
Qwen-32B	caramellic	clear	rev.	O	14	14	0	0	0
Qwen-32B	caramellic	ambiguous	rev.	11	26	1	0	25	0
Qwen-32B	caramellic	ambiguous	rev.	O	26	26	0	0	0
Qwen-32B	meaty	clear	rev.	11	14	6	0	8	0
Qwen-32B	meaty	clear	rev.	O	14	14	0	0	0
Qwen-32B	meaty	ambiguous	rev.	11	26	6	0	20	0

Continued on next page

Table 21 (continued)

Model	Query	Scope	Prompt	Cond.	N	C	W	T	I
Qwen-32B	meaty	ambiguous	rev.	O	26	26	0	0	0

J RETENTION AND NET GAIN ACROSS FIXED INTERFACES

The same accounting distinguishes complete retention from positive incremental value. The fixed-subset rows are recomputed from saved full-information responses; the full-cohort rows reproduce frozen summaries. No models were rerun. Table 22 holds the human pairs fixed across the original and revised prompts. Table 23 gives the original-prompt full-cohort results separately. The 8B configuration uses only the fixed subset and therefore has no full-cohort row. These are descriptive counts conditional on the fixed human references; no new confidence intervals are added.

R_{keep} reports retained rule-correct presentations; R_{stable} reports retained rule-correct pairs in both orders. S_{corr} is the number of explicitly rule-wrong pairs made correct in both orders over all explicitly rule-wrong pairs. Rule ties are excluded from that denominator. Δ_{rule} is the net correctness difference in Eq. 3, shown as an exact signed count difference divided by all presentations. Exact agreement counts identical A/B/`TIE` outputs; agreement on abstention is included. A zero opportunity denominator is NA, never $0/0$.

Under the revised prompt, both Phi-mini and Qwen-4B exactly match the score rule on all 36 clear floral presentations. They preserve all 28 rule-correct presentations and all 14 rule-correct pairs, with zero net gain. This is complete decision retention on this subset, without added correctness. It is an observed counterpoint to the initial released-floral cancellation, rather than evidence that retention must fail. Exact behavioral agreement does not identify the internal computation used by a model.

A favorable ambiguous-reference result also remains in the record: revised Phi-mini stably corrects one of four explicitly rule-wrong meaty pairs. This outcome is defined relative to the frozen panel-mean ordering; the ambiguous reference does not establish a certain perceptual difference. The absence of stable correction in the displayed clear-subset rows must not be generalized to every archived subset.

The accompanying machine-readable summary additionally includes every ambiguous subset, initial released-floral direct condition, and original released-label full-information factorial condition. It reports gains separately from explicit rule errors and rule ties, and losses separately into wrong choices, abstentions, and invalid outputs. The fixed original rows are historical subsets of the factorial records; displaying them again does not add independent observations or new generations. The reused local-model rows in the 32B comparison are excluded from duplicate reporting.

Table 22: Retention, stable correction, and net correctness on the fixed clear human subset. Floral/roasted/caramellic/meaty have 18/19/7/7 pairs. All rows use full prediction information. Orig. and rev. denote the original and revised prompt packages; 32B was evaluated only with the revised package.

Model	Query	Prompt	R_{keep}	R_{stable}	S_{corr}	Δ_{rule}	Agree
Phi-mini	floral	orig.	15/28	1/14	0/4	$-10/36$	19/36
Phi-mini	floral	rev.	28/28	14/14	0/4	$0/36$	36/36
Qwen-4B	floral	orig.	15/28	1/14	0/4	$-10/36$	20/36
Qwen-4B	floral	rev.	28/28	14/14	0/4	$0/36$	36/36
Qwen-8B	floral	orig.	0/28	0/14	0/4	$-28/36$	0/36
Qwen-8B	floral	rev.	0/28	0/14	0/4	$-28/36$	0/36
Qwen-32B	floral	rev.	18/28	7/14	0/4	$-10/36$	22/36
Phi-mini	roasted	orig.	17/32	2/16	0/3	$-14/38$	21/38
Phi-mini	roasted	rev.	29/32	13/16	0/3	$-3/38$	35/38
Qwen-4B	roasted	orig.	16/32	0/16	0/3	$-13/38$	19/38
Qwen-4B	roasted	rev.	20/32	4/16	0/3	$-9/38$	23/38

Continued on next page

Table 22 (continued)

Model	Query	Prompt	R_{keep}	R_{stable}	S_{corr}	Δ_{rule}	Agree
Qwen-8B	roasted	orig.	0/32	0/16	0/3	$-32/38$	0/38
Qwen-8B	roasted	rev.	0/32	0/16	0/3	$-32/38$	0/38
Qwen-32B	roasted	rev.	5/32	0/16	0/3	$-27/38$	5/38
Phi-mini	caramellic	orig.	2/6	0/3	0/3	$-3/14$	7/14
Phi-mini	caramellic	rev.	5/6	2/3	0/3	$0/14$	12/14
Qwen-4B	caramellic	orig.	3/6	0/3	0/3	$0/14$	8/14
Qwen-4B	caramellic	rev.	5/6	2/3	0/3	$+1/14$	11/14
Qwen-8B	caramellic	orig.	0/6	0/3	0/3	$-6/14$	2/14
Qwen-8B	caramellic	rev.	0/6	0/3	0/3	$-6/14$	2/14
Qwen-32B	caramellic	rev.	0/6	0/3	0/3	$-6/14$	2/14
Phi-mini	meaty	orig.	6/12	0/6	NA	$-6/14$	8/14
Phi-mini	meaty	rev.	12/12	6/6	NA	$0/14$	14/14
Qwen-4B	meaty	orig.	6/12	0/6	NA	$-6/14$	8/14
Qwen-4B	meaty	rev.	9/12	3/6	NA	$-3/14$	11/14
Qwen-8B	meaty	orig.	0/12	0/6	NA	$-12/14$	2/14
Qwen-8B	meaty	rev.	0/12	0/6	NA	$-12/14$	2/14
Qwen-32B	meaty	rev.	6/12	1/6	NA	$-6/14$	8/14

Table 23: The same full-information accounting on the original full clear human cohort: 34/37/15/15 pairs for floral/roasted/caramellic/meaty. These rows and the fixed-subset rows describe different sample scopes.

Model	Query	R_{keep}	R_{stable}	S_{corr}	Δ_{rule}	Agree
Phi-mini	floral	24/46	2/23	0/11	$-16/68$	35/68
Qwen-4B	floral	24/46	1/23	0/11	$-12/68$	36/68
Phi-mini	roasted	31/58	6/29	0/7	$-24/74$	40/74
Qwen-4B	roasted	29/58	0/29	0/7	$-22/74$	38/74
Phi-mini	caramellic	6/16	0/8	0/6	$-9/30$	13/30
Qwen-4B	caramellic	8/16	0/8	0/6	$-2/30$	16/30
Phi-mini	meaty	15/24	3/12	0/2	$-8/30$	19/30
Qwen-4B	meaty	13/24	1/12	0/2	$-9/30$	17/30

K Separating Option Labels, Display Order, and Aggregation

This follow-up separates the identity attached to an option label from its display position. It uses previously examined released-annotation pairs, not a new confirmation sample. The inputs, parser, aggregation rule, and analysis were frozen before the new generations. LF with Qwen3-4B was the designated primary comparison; the other five panels were specified secondary comparisons. The results remain exploratory.

Scope and encoding. LF uses the initial released-label floral D messages and their `model_score` field. FF uses the same floral pairs and displayed numbers in the factorial stage's original `ID1_SD1` interface, with a `model_mean` field. FO uses that factorial interface for odorless. Each stratum contains 101 pairs and 202 distinct molecular connectivity identities. LF and FF share all pairs; floral and odorless share 74 molecules. Thus neither the two floral interfaces nor the three pooled strata constitute independent replications. The reference is released descriptor membership: an unlisted descriptor is not an experimentally established sensory absence. These controls include neither the human-rating cohort nor nutty.

For each physical pair (m_1, m_2), Table 24 crosses two label assignments with two block orders. Each complete molecular block, including its structure and numerical evidence, moves with its physical molecule. No evidence is reassigned between molecules. E1 and E4 reproduce the two

original directions; E2 and E3 put a block labeled B first. All choices are mapped back to physical identity before comparison or voting.

Table 24: The four encodings. Entries show physical molecule and displayed label.

Encoding	First block	Second block	Relation to original directions
E1	m_1: A	m_2: B	Original forward
E2	m_1: B	m_2: A	Label exchange at fixed order
E3	m_2: B	m_1: A	Order exchange at fixed labels
E4	m_2: A	m_1: B	Original reversed

Parsing and numerical control. The parser accepts only the first nonempty line: A, B, or TIE, case-insensitively, with at most one of the frozen terminal punctuation marks. Later explanation cannot repair an invalid first line. Scientific `TIE` means abstention. The numerical control removes the scientific task and asks for the larger of the same displayed three-decimal values under the corresponding label/order mapping; there, `TIE` is correct exactly when the displayed numbers are equal. LF and FF share their numerical-control outputs. This control measures execution of this particular interface, not general arithmetic competence. Displayed ties are never broken with hidden precision or reference labels. FO has one displayed-tie pair; both models answer its four numerical presentations correctly with `TIE`.

Estimands. Each panel contains 404 presentations of 101 pairs. Presentation accuracy counts correct choices over all 404 outputs; scientific abstentions and invalid outputs receive no correctness. Answered risk uses wrong A/B choices divided by all A/B choices. Let S contain the 78 pairs on which the displayed rule is correct. We report R_2 and R_4, the fractions of S answered correctly in both E1/E4 and all four encodings, respectively. These conditional retention rates differ from correct-pair fractions over all 101 pairs. In these panels the numerators happen to coincide, but the denominators must not be exchanged. Applying a stricter all-four criterion does not itself estimate a causal intervention effect.

Table 25: All six panels. Scientific correct/wrong/`TIE` counts sum to 404; invalid and missing outputs are zero. R_2 and R_4 have denominator 78. Numerical correctness has denominator 404 and includes correct numerical ties. Phi denotes Phi-4-mini; Q4 denotes Qwen3-4B-Instruct-2507.

Stratum	Model	Correct	Wrong	`TIE`	R_2 / R_4	Numerical correct
LF	Phi	243	161	0	52/78; 0/78	234/404
LF	Q4	317	87	0	67/78; 67/78	401/404
FF	Phi	253	151	0	73/78; 0/78	234/404
FF	Q4	316	88	0	78/78; 78/78	401/404
FO	Phi	203	195	6	1/78; 0/78	235/404
FO	Q4	230	170	4	3/78; 3/78	389/404

Fixed vote and comparison rules. The only aggregate selects a physical molecule if at least three of its four outputs select that molecule; otherwise it abstains. Raw `TIE` and invalid outputs cast no vote. This requires four model observations per pair, before exact reuse. We compare the vote with the displayed rule on all pairs, the rule on exactly the vote-selected pairs, and an independent gap selector at the same answered count. For the last comparison, within displayed-score-nontied pairs, the selector retains the K largest absolute displayed gaps, where K is the vote's answered count. Gap ties are resolved by the frozen hash of seed 20260920 and pair identity. The selector never uses correctness labels, and the gap is not a calibrated confidence. Equal coverage need not select the same molecules. Zero answered denominators are reported as NA.

What the comparisons show. LF/Q4 is correct on 317/404 presentations, versus 312/404 for the displayed rule: $+1.24$ points, with interval $[-1.49, 3.96]$. Its 16 favorable presentation-level corrections coexist with 11 lost correct choices. All-four retention is 67/78; fixed-vote retention is 78/78. These are different criteria, and the vote corrects none of the 23 rule errors. Its overall correctness is 78/101, not 100%; its answered accuracy is 78/98. The observed risk is 20/98 versus

Table 26: Retention and correction use different opportunities. Presentation columns count all four encodings; vote columns count physical pairs. The rule is correct on 78 pairs in every stratum, wrong on 23 floral pairs and 22 odorless pairs, and tied on one odorless pair. Neither model corrects the FO rule-tie pair. "Lost" includes wrong aggregate choices and abstentions on rule-correct pairs.

		Per presentation		Fixed vote		
Stratum	Model	Retained	Corrected	Retained	Corrected	Lost
LF	Phi	208/312	35/92	52/78	0/23	26/78
LF	Q4	301/312	16/92	78/78	0/23	0/78
FF	Phi	229/312	24/92	73/78	0/23	5/78
FF	Q4	312/312	4/92	78/78	0/23	0/78
FO	Phi	154/312	49/88	1/78	6/22	77/78
FO	Q4	204/312	26/88	45/78	0/22	33/78

Table 27: Fixed-vote outcomes and independently gap-selected rule risk. Correct (C), wrong (W), and abstained (A) counts sum to 101. Both risks use the shown answered denominator. Risk differences are vote minus gap, in percentage points, with exploratory 95% intervals. FO matching excludes its single displayed tie, on which both votes abstain; its all-pair coverage remains answers/101.

Stratum	Model	C/W/A	Vote risk	Gap risk	Difference	95% interval
LF	Phi	52/11/38	11/63	9/63	$+3.17$	$[-4.69, 13.64]$
LF	Q4	78/20/3	20/98	21/98	-1.02	$[-2.08, 0.00]$
FF	Phi	73/22/6	22/95	19/95	$+3.16$	$[0.00, 6.59]$
FF	Q4	78/21/2	21/99	21/99	0.00	$[0.00, 0.00]$
FO	Phi	7/4/90	4/11	1/11	$+27.27$	$[-14.29, 62.50]$
FO	Q4	45/18/38	18/63	10/63	$+12.70$	$[4.55, 25.00]$

21/98: one fewer error, with a difference interval that includes zero. Each vote requires four model observations; gap selection uses existing scores. This is a matched-coverage descriptive comparison, not a cost-matched deployment comparison or evidence of voting superiority.

FF/Q4 follows the displayed rule in all four encodings on 99/101 pairs and always outputs label A on the other two. The vote again retains 78/78 rule-correct pairs without correcting a rule error. Its observed risk equality with the matched gap selector is not a statistical equivalence result. LF and FF reveal an interface difference on shared material, not generalization to an independent task.

FO/Phi supplies a different local benefit: on the vote's 11 selected pairs, voting is correct on 7, while the displayed rule is correct on 4. The difference comprises six corrected rule errors and three lost correct decisions. Yet independent gap selection is correct on 10/11, and the vote abstains on 90/101 pairs overall. These same-subset and same-coverage comparisons answer different questions. Across all six panels, no rule-wrong pair is corrected in every encoding; this does *not* mean that every vote has zero corrections.

Observable patterns and limits of interpretation. On FO's 100 nontied pairs, Phi always selects the first displayed molecule on 84, always selects A on 3, and has mixed behavior on 13. Q4 has the corresponding counts 29, 8, and 60, with 3 additional pairs always following the displayed rule. These are patterns in choices, not identified internal mechanisms. Phi's numerical correctness is only 234/404 for floral and 235/404 for odorless (approximately 58%). It therefore cannot establish a cross-model replication in which basic execution of this interface has been ruled out. Q4's 401/404 and 389/404 controls are stronger, but also imperfect. Mixed behavior and unexplained residuals do not by themselves establish a novel mechanism, and these annotation-task controls cannot explain the separate human-rating results by attribution alone.

Raw scientific abstentions and vote abstentions are also distinct. In table order, vote nonselection comprises 38 conflicts; 3 conflicts; 6 conflicts; 2 conflicts; 89 conflicts plus one insufficient-vote case containing raw `TIE`; and 37 conflicts plus one all-`TIE` case. The frozen classification gives physical-choice conflict priority and can assign a case containing raw `TIE` to that category. This accounting is not a causal decomposition. No invalid output contributes to these results.

Uncertainty and implementation. The archived intervals use 20,000 bootstrap draws with seed 20260920 and percentile endpoints. Each draw retains all four observations of a pair and resamples molecular-sharing connected components; these are 101 single-pair components within each stratum. Ratios, K, and gap selection are recomputed in each draw. The intervals are conditional on the fixed cohorts, outputs, and interface configurations; they do not include retraining, new sources, or model selection. They are unadjusted exploratory intervals, not full-table significance tests. Connectivity disjointness does not establish independent data collection. Zero-width intervals at boundary outcomes provide no population guarantee. Historical intervals retain their original scopes and are not replaced by these intervals.

Both local configurations use bfloat16, eager attention, batch size one, greedy decoding, at most 64 output tokens, and archived tokenizer/chat templates. The checkpoints are Phi-4-mini-instruct revision `cfbefacb99257ffa30c83adab238a50856ac3083` and Qwen3-4B-Instruct-2507 revision `cdbee75f17c01a7cc42f958dc650907174af0554`. Qwen uses its non-thinking template; Phi has no thinking toggle. Runtime fingerprints and decoding settings are archived. There were no technical failures, retries, missing outputs, invalid scientific outputs, or truncations. Generation and historical-reuse accounting are given in Appendix G.

K.1 Matched interfaces and encoding-specific outcomes

Table 28 places aggregate correctness and stable retention side by side for Qwen3-4B. All 101 pairs have identical physical identities, structures, released references, displayed means, and displayed dispersion in LF and FF. Their E1/E4 accuracy is exactly equal, while the retained rule-correct sets differ. The full interface messages and execution configuration are given in Appendix K.2; this comparison does not isolate a single textual cause.

Table 28: Matched Qwen3-4B interface comparison on the same 101 floral pairs, displayed scores, and references. R_2 and R_4 count rule-correct pairs retained in both original encodings or all four encodings, respectively. Equal aggregate correctness under E1/E4 coexists with different retention. LF and FF differ as complete interface packages, not as a single-factor ablation.

Interface	E1/E4 correct	R_2	E1–E4 correct	R_4
LF	158/202	67/78	317/404	67/78
FF	158/202	78/78	316/404	78/78

Table 29 reports every encoding separately. For LF/Qwen, scientific correct counts are 91, 78, 81, and 67; for FF/Qwen they are 80, 78, 80, and 78. Their shared numerical control yields 101, 101, 98, and 101 correct answers. On FO/Qwen, scientific correct counts are 99, 75, 53, and 3, whereas numerical total correctness is 101, 99, 88, and 101. The reference targets differ between scientific and numerical tasks; the latter asks only for the larger displayed number, with `TIE` correct on equality.

Table 29: Outcomes by encoding, with 101 observations per cell. Each tuple is C/W/TIE/invalid. Scientific C/W is relative to released annotations; numerical-control C/W is relative to the displayed numbers. To keep categories disjoint, numerical C counts correct A/B outputs only. Every numerical TIE shown here is correct on an equal-number input, so total numerical correctness is C+TIE. LF and FF share the same numerical-control calls.

Panel	Model	Encoding	Scientific C/W/T/I	Numerical C/W/T/I
LF	Phi	E1	52/49/0/0	44/57/0/0
LF	Phi	E2	101/0/0/0	78/23/0/0
LF	Phi	E3	0/101/0/0	23/78/0/0
LF	Phi	E4	90/11/0/0	89/12/0/0
LF	Qwen3-4B	E1	91/10/0/0	101/0/0/0
LF	Qwen3-4B	E2	78/23/0/0	101/0/0/0
LF	Qwen3-4B	E3	81/20/0/0	98/3/0/0
LF	Qwen3-4B	E4	67/34/0/0	101/0/0/0

Continued on next page

Table 29 (continued)

Panel	Model	Encoding	Scientific C/W/T/I	Numerical C/W/T/I
FF	Phi	E1	75/26/0/0	44/57/0/0
FF	Phi	E2	101/0/0/0	78/23/0/0
FF	Phi	E3	0/101/0/0	23/78/0/0
FF	Phi	E4	77/24/0/0	89/12/0/0
FF	Qwen3-4B	E1	80/21/0/0	101/0/0/0
FF	Qwen3-4B	E2	78/23/0/0	101/0/0/0
FF	Qwen3-4B	E3	80/21/0/0	98/3/0/0
FF	Qwen3-4B	E4	78/23/0/0	101/0/0/0
FO	Phi	E1	100/0/1/0	45/55/1/0
FO	Phi	E2	93/7/1/0	78/22/1/0
FO	Phi	E3	9/89/3/0	22/78/1/0
FO	Phi	E4	1/99/1/0	86/14/1/0
FO	Qwen3-4B	E1	99/1/1/0	100/0/1/0
FO	Qwen3-4B	E2	75/25/1/0	98/2/1/0
FO	Qwen3-4B	E3	53/47/1/0	87/13/1/0
FO	Qwen3-4B	E4	3/97/1/0	100/0/1/0

For Phi, E2 and E3 always select label B on floral. B is first in both encodings but refers to different physical molecules, giving 101/101 and 0/101 scientific correctness, respectively. This observable response pattern accompanies poor numerical-control performance and therefore does not provide a capability-controlled cross-model mechanism explanation.

All 11 rule-correct LF/Qwen pairs recovered by voting have the same physical-choice pattern (Table 30). Their correct molecule m_1 is missed only in E4, where it is labeled B and placed second. It is selected in E2 when labeled B but placed first, and in E3 when placed second but labeled A. These are observed label/order contingencies, not evidence identifying an internal interaction mechanism. Other errors and all six panels remain in Table 29.

Table 30: The shared physical-choice pattern of all 11 rule-correct LF/Qwen3-4B pairs recovered by voting. The correct physical molecule is m_1. Only E4 combines label B with second position for m_1. This is an observed encoding-specific pattern, not an identified internal mechanism.

	E1	E2	E3	E4	Vote
First option	A:m_1	B:m_1	B:m_2	A:m_2	–
Second option	B:m_2	A:m_2	A:m_1	B:m_1	–
Answer label	A	B	A	A	–
Physical choice	m_1	m_1	m_1	m_2	m_1

The recovered pairs have frozen floral indices 006, 007, 013, 018, 027, 042, 055, 059, 082, 097, and 099. Complete pair identifiers and the four physical choices are retained in the accompanying ledger. Each pair's historical E1/E4 decisions match its regenerated decisions individually.

K.2 Complete LF and FF interfaces and execution configuration

LF and FF share all 101 floral pairs, reference labels, SMILES strings, displayed means, and displayed SDs. The following system and user messages reproduce one common E1 pair in full. Their scientific content and field names are unchanged from the archive; paragraph breaks below follow the input line breaks. E2–E4 move the complete molecule–evidence blocks and relabel A/B according to Table 24.

LF places the database-association task, score interpretation, annotation boundary, and reference-label instructions in the system message. FF uses a general comparison system message, places the database-label task in user context, changes `model_score` to `model_mean`, and revises the question and abstention wording. The system also differs in its explanation of predictions, five-member SD, and supplied-reference controls. These changes form a complete interface package; no result isolates a field-name or sentence effect.

LF system message.

> Compare molecular association with an odor descriptor in a database-label task, not perceived odor intensity. Use the displayed molecule identities and visible information. Model scores are predictions: a higher score expresses greater predicted association with its named descriptor. Ensemble standard deviation is dispersion across model members, not a confidence interval or a calibration guarantee. When reference labels are shown, they are an oracle positive control: 1 means the descriptor is listed in the database and 0 means it is not listed; an unlisted annotation is not a measured negative. Judge the displayed molecules using the available information.

LF user message.

> Query descriptor: ”floral”
> Molecule A SMILES: CC(CC=O)C(C)c1ccccc1
> Evidence for molecule A:
> descriptor=floral; model_score=0.048; ensemble_sd=0.023
> Molecule B SMILES: CCCCCC(O)(O)CC
> Evidence for molecule B:
> descriptor=floral; model_score=0.062; ensemble_sd=0.048
> Question: Which molecule is more associated with the query descriptor ”floral” in the database-label task?
> Output contract: On the first line, output exactly A, B, or TIE. Do not output an explanation. TIE means that you do not select A or B; it does not assert equal true odor association.

FF system message.

> You compare two molecules using the visible information. A model mean is a predicted descriptor association, not a calibrated probability or a human rating. Ensemble SD is dispersion across five fitted models, not an error probability, confidence interval, or guarantee. Use only the available information. You may decline to select a molecule by answering TIE. When reference measurements are provided, they are a label-assisted positive control, not model predictions.

FF user message.

> Query descriptor: ”floral”
> Task: compare released database descriptor associations. A listed annotation is positive; an unlisted annotation is not a measured negative. This is not a comparison of perceived odor intensity.
> Molecule A SMILES: CC(CC=O)C(C)c1ccccc1
> Evidence for molecule A:
> descriptor=floral; model_mean=0.048; ensemble_sd=0.023
> Molecule B SMILES: CCCCCC(O)(O)CC
> Evidence for molecule B:
> descriptor=floral; model_mean=0.062; ensemble_sd=0.048
> Question: Which molecule is more associated with the database descriptor ”floral”?
> Output contract: On the first line, output exactly A, B, or TIE. Do not output an explanation. TIE means that you decline to select A or B; it does not assert that the true ratings or associations are equal.

Execution configuration and historical correspondence. Within each model, regenerated LF and the four-encoding FF analysis use one frozen runtime configuration: BF16, eager attention, batch size one, generation seed 20260916, greedy decoding (`do_sample=false`, one beam), and at most 64 generated tokens. The retained revisions are `cdbee75f17c01a7cc42f958dc650907174af0554` for Qwen3-4B-Instruct-2507 and `cfbefacb99257ffa30c83adab238a50856ac3083` for Phi-4-mini. Runtime hashes also fix the model files, chat template, parser, and recorded software configuration.

FF E1/E4 uses 202 historical outputs per model whose archived reuse checks passed for messages, tokenization, model files, source revision, decoding, seed, parser, environment, and raw runtime

hash; FF E2/E3 was generated in the follow-up. All LF encodings were generated in that follow-up because the historical LF runtime-lock hash could not be reconciled. Direct comparison nevertheless finds identical physical choices, raw text, and generated tokens on all 202 shared LF presentations per model, with zero changed pairs. The 11 Qwen and 26 Phi rule-correct pairs not retained in both original orders are the same pair identities in both records. This establishes output-level correspondence for the historical ledger, not reconstruction of an unverified historical runtime.

L Audit Protocol and Anonymous Evidence Map

The anonymous supplement, `CAVeR_Anonymous_Audit.zip`, connects Figure 1 to saved inputs, responses, and decisions. This appendix specifies the reusable protocol and maps the principal results. The full machine-readable map, `materials_to_paper.csv`, uses LaTeX figure and table labels.

L.1 Specifying a new audit instance

The protocol requires a fixed predictor, identifiable candidates, a displayed-score rule, and a stated reference, without requiring olfactory descriptors or molecular representations. Applying it elsewhere requires a new empirical evaluation; the present effects do not establish generality.

1. **Fix the comparison.** Freeze candidate identities, predictor, displayed precision, numerical decision rule, and correctness reference. Separate observed targets from proxy predictions; retain all ties and uncertain reference cases.
2. **Define matched interventions.** Reassign score blocks at fixed presentation to separate behavioral response from assignment advantage. Exchange labels and positions with evidence attached to each candidate. Save complete original and revised messages: a joint wording revision identifies a package contrast, not a sentence-level effect.
3. **Preserve decision identity.** Store raw responses and parsing outcomes; map labels back to physical candidates. Distinguish abstention from invalid output. Link observations to pair, query, configuration, prompt, and source record.
4. **Account for gains and losses.** Separate retention on rule-correct opportunities, corrections on explicit rule errors, and gains from rule ties. With repeated presentations, report presentation-level and stable pair-level outcomes. Recovering a correct decision need not improve on the rule.
5. **Compare selective decisions.** At each observed model coverage, select the same answer count with a fixed numerical selector. Report membership and risk; equal coverage need not match selected cases or computational cost.
6. **Delimit uncertainty and reuse.** Identify sampling units and shared-candidate dependencies; report effect intervals where supported. Distinguish generation events from unique questions. Record adaptive follow-ups, configuration gaps, and supplied-target reading-control scope.

These steps organize established comparisons into a decision-level audit. They are not a new correctness metric, a calibration guarantee, or a mechanism-identification procedure.

L.2 From saved outputs to reported results

Run `python run_reanalysis.py` offline with Python 3.10 or later; no model weights, GPU, credentials, or network connection are needed. The runner checks file hashes, parses saved outputs, reconstructs displayed-score rules and physical choices, and compares regenerated CSV rows with frozen exports. It writes `recomputed/verification.json` and terminates on a mismatch. Plot regeneration is separate and requires PyMuPDF and NumPy.

Raw records and joins. Initial released-floral files `historical/saved_inputs.jsonl` and `historical/saved_outputs.jsonl` join by model and input key. Label–position records join `gate/inputs/observations.jsonl` to `gate/outputs/<model>`

/<call_id>.json; analysis uses may share an output. Human-rating joins are specified in human_audit/README.md. Displayed-score ledgers preserve the numerical strings shown to the model, without substituting hidden precision. Pair ledgers expose individual gains, losses, abstentions, and stable outcomes.

Table 31: Entry points for the main claims. Paths are relative to the anonymous supplement; the full map includes intermediate records and appendix results.

Paper result	Saved records, intermediate ledgers, and analysis
Figure 1: estimands	figures/build_figures.py renders the definitions in Section 3. This conceptual figure contains no estimated effects or synthetic observations.
Table 1: response and assignment	assignment_response/recompute_assignment_response.py parses the historical records. assignment_response/assignment_presentation_ledger.csv and assignment_response/physical_state_transitions.csv expose matched states; assignment_response/assignment_response_summary.csv contains the point estimates.
Table 2: retained and lost pairs	scripts/trace_decisions.py reconstructs the displayed rule across both orders. expected_traces/historical_pair_transition_ledger.csv lists pair outcomes; expected_traces/historical_retention_summary.csv reproduces the retention counts.
Figure 2 and Table 3: shared pairs, encoding, and voting	encoding_controls/analyze_existing_encodings.py reparses scientific and numerical observations. scripts/trace_decisions.py independently checks fixed voting and gap selection. expected_traces/gate_pair_transition_ledger.csv records four responses and selection membership per pair.
Figure 3 and Table 4: human-rating contrasts	human_audit/recompute.py checks saved human-task responses and rendered evidence. human_audit/expected/wording_transition_ledger.csv traces prompt-revision decisions; human_audit/expected/pair_retention_ledger.csv separates rule-relative retention and correction; human_audit/expected/reference_selector_ledger.csv records matched-coverage selection. figures/build_figures.py renders the plotting data.

Recomputation boundary. Replay covers point estimates and decision identities for the initial released-label floral, four-encoding, matched human-interface, and 32B records. Other full-cohort factorial results in frozen_full_cohort/ are frozen aggregates without independent raw-response verification. The runner does not retrain the predictor, regenerate responses, or reproduce sensory measurements. Historical confidence intervals remain frozen with their original scope; replay neither resamples them nor independently validates every interval. Generation counts are not independent-task counts. Upstream release provenance and the hosted model's exact revision and precision remain unresolved.

Integrity and anonymity. MANIFEST.csv records included-file hashes; source_manifest.csv records source hashes and transformations. Machine-specific cache/interpreter paths are omitted; scientific configuration and retained messages/responses are preserved. Runtime fingerprints refer to original metadata, not the redacted export. ANONYMIZATION.md and DATA_SOURCES.md document these distinctions and attribution. Packaging neither repairs provenance nor licenses third-party data.