WHEN FRAMING CHANGES THE SCORE: A COMPARATIVE STUDY OF LLM JUDGE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models used as judges can change their ratings in response to an answer's framing as well as its content. We study how these sources of score change relate internally and what their relationship implies for controlling ratings. Behavioral measurements across seven judges and nine benchmarks show widespread but heterogeneous framing responses. A controlled comparison in Llama-3.1-8B-Instruct and Qwen3-4B-Instruct-2507 matches reviewer-consensus framing to known final-answer errors on the original rating and score decrease. Both sources have concentrated activation changes, and their mean directions align closely at the final blocks, with cosines of 0.880 and 0.901. Interventions clarify how these directions influence scoring. A framing direction's components can differ in the magnitude and sign of their effects, and a component with little effect at its original magnitude can control ratings after independent scaling. Directions learned from unframed ratings also control scores, while ranking preservation varies across configurations. In Llama, moving each framed rating toward its unframed reference can narrow the score gap between clean and corrupted answers. Together, the source comparison and interventions connect framing responses to broader scoring behavior: distinct input changes approach similar late-layer mean directions, and control along these directions extends beyond the targeted rating.

1 INTRODUCTION

Large language models serve as evaluators of generated text, supporting model comparison and the assessment of conversational responses, summaries, and other outputs (Zheng et al., 2023; Liu et al., 2023; Gu et al., 2024). Their ratings depend on both the candidate answer and the context in which it is presented. Authority cues and prior-evaluation metadata, for example, can change a judgment while leaving the answer itself fixed (Chen et al., 2024; Kapetanovic et al., 2026). Such framing responses raise a question about the evaluator's internal computation: how does a change in evaluation context relate to the representations used to assign a score?

An unfavorable description of an unchanged answer and an error introduced into that answer can both lower a judge's rating. The same score change can therefore arise from different input sources. Comparing those sources at matched original scores and score decreases allows us to ask whether their internal changes also resemble one another. This comparison builds on two complementary lines of work. Internal rating methods show that hidden states and score-token probabilities contain useful evaluation signals (Lai et al., 2025; Girrbach et al., 2025), while activation steering tests how fitted directions affect model behavior (Turner et al., 2023; Rimsky et al., 2024). Together, they make it possible to study both the relationship between sources of rating change and the consequences of intervening along the resulting directions.

Our behavioral measurements across seven judges, nine benchmarks, and seven framing families establish a widespread but heterogeneous response. Unfavorable framing lowers the aggregate rating in every judge, with substantial differences by cue, domain, and prompt. We investigate the internal response in LLAMA-3.1-8B-INSTRUCT and QWEN3-4B-INSTRUCT-2507 using a shared, independent cohort of 1,200 answers. Reviewer-consensus and review-status metadata change the evaluation context, while incorrect final-answer edits change the answer content. Measuring each

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.

altered input against its own unframed (clean) counterpart makes these sources directly comparable. For consensus framing, the score-matched mean displacements align closely with those from answer errors in the later observed layers of both judges, reaching cosines of 0.880 and 0.901 at the final blocks. Concentrated variation around each source's own mean accompanies this alignment.

Interventions establish what these geometric relationships mean for the judge's evaluations. Using references learned from clean ratings, we separate framing directions into parallel and residual components and test them at their original magnitudes and after independent scaling. A component's share of geometric energy does not determine the magnitude or sign of its score effect. Generic score directions also change ratings, making the preservation of answer rankings an important part of the comparison. The same issue arises when the goal is to restore a framed rating: Llama's paired recovery curves show movement toward the clean reference alongside changed rankings and a narrower clean/error score gap. These results place framing-derived directions within the judge's broader scoring behavior. Their interpretation depends on both the response used to estimate them and the other evaluations they change.

2 Related Work

LLM-based evaluation studies both agreement with human judgments and sensitivity to evaluation context (Zheng et al., 2023; Liu et al., 2023; Gu et al., 2024). Position, self-preference, authority cues, and prior-evaluation metadata can affect these judgments (Wang et al., 2024; Panickssery et al., 2024; Chen et al., 2024; Kapetanovic et al., 2026). Alongside these behavioral studies, internal rating methods recover evaluation signals from the model itself: LAGER combines cross-layer score-token logits, while Latent Judges uses rating probabilities and activation probes (Lai et al., 2025; Girrbach et al., 2025). We connect the two perspectives by comparing the internal changes produced by framing and answer errors at matched scoring outcomes, with unframed evaluations providing a score-related reference.

Activation Addition and contrastive steering derive directions that alter a frozen model's behavior (Turner et al., 2023; Rimsky et al., 2024); directional interventions also support functional studies of refusal (Arditi et al., 2024). We use such interventions to test the consequences of the observed representation relationships. Applying the directions to framed, clean, and answer-corrupted inputs reveals how score control extends across the judge's evaluations. Appendix A develops the relationships to internal rating, steering, and linear concept erasure.

3 Study Design

We study a judge's response to changes in its evaluation input at two scales. A broad behavioral study characterizes framing responses across seven judges, seven framing families, and nine question benchmarks. A controlled study then compares framing with answer errors in two white-box judges, using a separate set of answers shared by both models. Holding the input transformations and comparison procedure fixed allows us to relate the resulting internal changes and test their effects on ratings.

3.1 Paired Inputs and Source Comparisons

The controlled study starts with 1,200 questions: 400 each from GSM8K, CommonsenseQA, and ARC-Challenge (Cobbe et al., 2021; Talmor et al., 2019; Clark et al., 2018). Phi-4 (Abdin et al., 2024) generates one candidate answer per question, and both Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-4B-Instruct-2507 (Qwen Team, 2025) evaluate the same answers. Each benchmark contributes 240/80/80 questions to training/development/testing, for 720/240/240 overall. All variants stay in the split of their parent question, so direction estimation, configuration selection, and final evaluation use disjoint questions. Both correct and incorrect answers enter the framing study. Section 4.1 and Appendix B describe the broad study's separate population.

An input $x_i = P(q_i, a_i)$ presents question q_i and answer a_i to the judge, which assigns a 1–10 rating $s(x_i)$. We call this unframed input *clean*; the term refers to the absence of added metadata rather than answer correctness. We append reviewer-consensus metadata (*Bandwagon*) or review-status metadata (*Refinement*), leaving the answer body unchanged. Bandwagon reports a proportion of

reviewers who rated the answer as high or low quality; Refinement states whether the answer has been reviewed. Favorable and unfavorable wording indicate the intended influence, and the paired ratings establish its actual effect. Template A is used throughout the study; an alternative wording B is reserved for test questions.

For metadata condition c, the resulting input x_i^c gives a paired score change and hidden-state displacement:

$$\Delta s_i^c = s(x_i^c) - s(x_i), \qquad \Delta h_{i,l}^c = h_l(x_i^c) - h_l(x_i). \tag{1}$$

Here $h_l(x)$ is the raw output of decoder block l at the final input token. Following the complete evaluation input and preceding rating generation, this position provides a common observation point for contextual and score-related changes. Both judges receive a non-CoT scoring prompt and decode greedily with an eight-token cap. We extract the primary rating from the first completed integer line and separately measure whether the entire response contains only that integer (Appendix D.4).

Two controls provide comparisons with the evaluative metadata. Neutral metadata retains the family's header and bracketed format but substitutes an archive code for the evaluative statement. Its complete native-chat input has the same token count as the corresponding unfavorable input under that judge's tokenizer. The content control instead introduces a known answer error: when an answer's final field matches the benchmark reference, we replace that field with an incorrect number or option. The surrounding text remains unchanged, including any reasoning that now disagrees with the final field. This edit supplies a substantive source of score change against which to compare framing.

We match the framing and content arms on benchmark, clean rating, and score decrease to compare internal changes at the same observed scoring outcomes. Matching is performed separately within each judge and split, giving the two arms identical distributions of these quantities. A match can join different questions; each displacement nevertheless compares an altered input with its own clean input. Training matches estimate source-specific mean directions, while test matches compare geometry on new questions. Matching equalizes the rating response without imposing similarity on the hidden states, leaving their relationship to be measured. Appendices D.2 and E.3 provide the literal transformations, parent-answer eligibility, and exact matching procedure.

3.2 Framing Directions and Clean-Score References

Metadata can change a hidden state without changing the reported rating. We focus direction estimation on realized responses by using training pairs whose ratings move in the cue's intended direction. With $\epsilon_c = +1$ for favorable and -1 for unfavorable wording, this effective set is $\mathcal{E}_c = \{i : \epsilon_c \Delta s_i^c > 0\}$. Its mean displacement defines the framing direction:

$$\bar{\Delta} h_l^c = \frac{1}{|\mathcal{E}_c|} \sum_{i \in \mathcal{E}_c} \Delta h_{i,l}^c, \qquad v_l^c = \frac{\bar{\Delta} h_l^c}{\|\bar{\Delta} h_l^c\|_2}. \tag{2}$$

The direction summarizes the average internal change accompanying the intended rating response, while response counts establish how often that response occurs. We test the direction's effect on every clean test input, including questions whose framed ratings stayed unchanged or moved oppositely. We also compare the response groups within each cue along the training-fitted directions (Appendix E.2).

We relate the framing direction to ordinary score variation using clean evaluations alone. Scalar ridge regression predicts the judge's rating from its hidden state. A high–low contrast instead compares mean hidden states between highly and poorly rated answers, balancing the two groups within each contributing benchmark. The references capture different relationships with the score: one predicts ratings from individual states, while the other follows the difference between rating-group means. Both are fitted on training questions, with development data selecting the ridge regularization.

Let w_l be the scalar-regression coefficient and m_l the high–low difference, both expressed in raw activation coordinates. We form their joint reference separately in each model:

$$\mathcal{U}_l = \mathrm{span}\{w_l, m_l\}, \qquad \Pi_l = Q_l Q_l^\top, \tag{3}$$

where the columns of Q_l are an orthonormal basis, giving a reference of rank at most two. Held-out prediction assesses how well the scalar readout reconstructs clean ratings, and the joint span locates framing directions relative to both score-related references. The following parallel and orthogonal components are defined with respect to this fitted span (Appendix E.1).

The framing direction separates into a parallel component $v_l^{c,\parallel} = \Pi_l v_l^c$ and a residual $v_l^{c,\perp} = (I - \Pi_l)v_l^c$. Their squared norms sum to one and quantify the shares of mean-direction energy inside and outside the reference. Variation across individual pairs requires a different measurement: we center each source's displacements and compute its covariance spectrum. Mean-direction cosine measures alignment between sources, whereas the spectrum measures concentration around each source's own mean. The appendix additionally gives total and centered residual energy, distinguishing how much of the displacement remains outside the reference from how that variation is distributed.

These geometric measurements prepare the functional test. Component size measures a share of the observed direction; intervention reveals how the judge responds to an addition along that component. Comparing original-magnitude and normalized components allows us to examine both their contribution to the full addition and their ability to control ratings when scaled independently.

3.3 Functional Evaluation

We add αv once to the final input-token state and let the remaining computation produce a rating. Geometry is observed at half-depth, three-quarter-depth, and the final decoder block; interventions use the three-quarter-depth block in each judge. Applying a framing-derived direction to clean inputs tests its effect without the metadata used to estimate it. In Llama, we also apply each intervention to framed and content-altered inputs to examine how the effect depends on the answer being evaluated.

The component comparison separates magnitude from orientation. ***Raw components*** keep their projected norms and use the full direction's selected coefficient α, measuring each contribution at its size in the full addition. ***Unit components*** are normalized separately and receive their own strength selection, measuring the score-control ability of each orientation under a common search budget. We evaluate the raw components individually as well as together because the judge's nonlinear response need not make their score effects additive.

For the two-model forward comparison, development data select the largest mean decrease in clean ratings subject to at least 93% extractable ratings and Spearman correlation $\rho_S \geq 0.85$ with clean scores. Every unit direction uses the same signed search grid, with relative strength r expressing the coefficient as a fraction of the median clean training-state norm. Alongside each unfavorable framing mean and its components, we test scalar-ridge and high–low directions, neutral means, matched-source means with sufficient support, and five independently searched random directions. We then evaluate the selected configurations on all clean test questions, measuring score shift, output validity, and ranking preservation together. Appendix E.4 gives the grid, selection rules, and the zero-intervention reference used when no candidate is feasible.

Recovery of a framed rating targets a different outcome: moving each answer toward its own clean score. We measure paired recovery as the reduction in absolute error to that score. In Llama, we evaluate this improvement alongside changes to clean and content-altered ratings; their clean-minus-content score gap measures separation of an answer from its corrupted counterpart. This recovery study uses a logit-augmented clean reference, defined separately from the common two-direction reference above. Development-selected configurations assess recovery at a chosen operating point. Descriptive curves over the complete strength grid show how paired error, ranking, and other evaluations vary together, without selecting an operating point from test results (Appendix K).

Controlled-study 95% intervals use benchmark-stratified question-cluster resampling, sharing question identities and draws across method contrasts. We also evaluate the same interventions under the strict whole-response scoring rule. Common-reference comparisons and response curves retain the originally valid reference population and penalize later invalid outputs, keeping the evaluated questions fixed as output validity changes. Complete-pair estimates and output coverage accompany these results. Appendix E.5 gives the endpoint definitions and uncertainty procedure; the broad behavioral study retains its separate descriptive aggregation.

4 Results

The behavioral study establishes how framing responses vary across evaluation settings. The controlled study then relates those responses to answer errors, clean-score directions, and the effects of intervention.

(a) Seven judges; $n = 2{,}700$

Judge	Baseline mean	Negative $\Delta\bar{s}$
LLAMA-3.1-8B	5.71	−0.853
QWEN3-14B	5.53	−0.935
GEMMA-3-12B	5.61	−0.870
GPT-4O-MINI	5.48	−0.685
LLAMA-3.3-70B	5.39	−0.610
DEEPSEEK-V3	5.42	−0.545
GPT-4.1	5.27	−0.360

(b) Seven cue families; $n = 4{,}500$

Cue family	Positive $\Delta\bar{s}$	Negative $\Delta\bar{s}$
Prestige	+0.01	−0.38
Verbosity	+0.16	−0.60
Bandwagon	−0.46	−1.56
Authority	+0.02	−0.79
Sentiment	+0.17	−0.07
Refinement	+0.77	−0.57
Diversity	−0.19	−0.94

Table 1: Broad framing responses under a strict grading rubric without CoT. (a) Seven judges on 2,700 nested-test questions each; negative shifts average all seven unfavorable cue families. (b) LLAMA-3.1-8B on the full pool of 4,500 questions; shifts are relative to the baseline mean of 5.84. Positive and negative denote assigned framing polarities. Values are descriptive mean changes in score points. Full statistics appear in Tables 14 and 15.

4.1 FRAMING RESPONSES ACROSS JUDGES AND DOMAINS

The broad study uses 4,500 question–answer pairs: 500 questions from each of nine benchmarks, with answers drawn from six generator models. Applying both polarities of seven cue families to each answer produces 31,500 variants per polarity. The seven-judge panel uses a nested test set of 2,700 questions, while the family and domain comparisons use the full pool. Appendix B specifies the generators, judges, transformations, and partitions.

Unfavorable framing lowers mean ratings in all seven judges, with aggregate decreases of 0.360–0.935 points (Table 1). The shared downward response combines markedly different family-level effects. On the full LLAMA-3.1-8B pool, Bandwagon− lowers scores by 1.56 points, while Sentiment− lowers them by only 0.07. Favorable framing is less consistent: Refinement+ raises scores by 0.77, but favorable Bandwagon and Diversity lower them by 0.46 and 0.19. The wording's intended influence therefore provides a starting condition for measurement, rather than determining the judge's response.

The family-level means also conceal differences across question domains (Figure 1). Bandwagon− decreases reach 2.46 points on CommonsenseQA and 2.13 on ARC-Challenge. Authority− instead has its largest decreases on PubMedQA and GPQA (1.25 and 1.18), while Diversity− reaches 2.03 on BBQ and 1.86 on SocialMaze. Sentiment− remains small across all nine domains. Consensus thus produces a broadly strong response, whereas the largest effects of other metadata depend more on the questions being evaluated. Table 13 gives the full domain comparisons for both polarities.

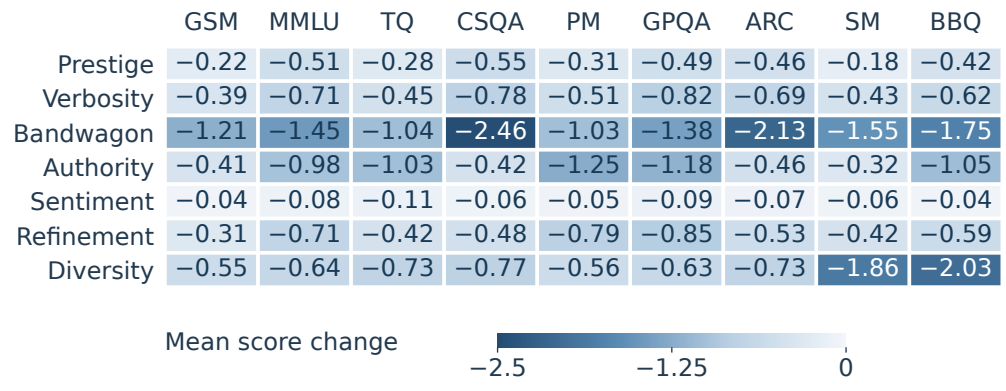


Figure 1: Unfavorable-cue mean score changes across nine benchmarks on Llama 3.1-8B. Each cell averages $s_{\text{cue}} - s_{\text{base}}$ over 500 pairs; darker blue denotes a larger decrease. TQ: TruthfulQA; PM: PubMedQA; SM: SocialMaze; ARC: ARC-Challenge. All values are point estimates.

Changing the evaluation prompt preserves the aggregate downward response but alters its magnitude. Across four GPT-4.1 configurations that vary the grading rubric and use of CoT, all-seven-family unfavorable means decrease by 0.350–0.855 points, while the corresponding favorable means remain near zero (Table 16). Together with the cue and domain comparisons, these results establish framing sensitivity across evaluation settings and show why a controlled internal comparison must specify the input change being studied.

Human comparisons examine whether the transformations retain the answer's substantive quality. Across six cue families, three annotators per pair evaluate 250 pairs per family for factual consistency, logical coherence, completeness, and overall quality relative to the original answer. Aggregate ratings satisfy the equivalence criterion on every dimension, using two one-sided tests with a 0.30 margin around the equivalent rating of 3 on a five-point scale (Appendix B.3). These comparisons support preservation of substantive quality for the sampled transformations under this rubric, complementing the judge scores with direct assessments of the transformed answers against their originals.

We next examine the internal changes accompanying these rating responses. The controlled study begins with all response outcomes and then estimates directions from the pairs that move in the cue's intended direction.

4.2 Different Metadata Produce Different Responses

On the same controlled-study questions, both judges respond much more frequently to unfavorable reviewer consensus than to a statement that the answer has not been reviewed.

Bandwagon− lowers 239 of 240 Llama test ratings and 234 of 240 Qwen ratings. Refinement− lowers only 24 and 14, leaving 208 and 222 unchanged, respectively (Figure 2); its remaining responses move upward despite the unfavorable wording. The training populations show the same contrast. Bandwagon and Refinement direction estimates use 715 versus 71 effective pairs in Llama and 700 versus 36 in Qwen. Thus, the two directions summarize responses that differ substantially in frequency within the same question population.

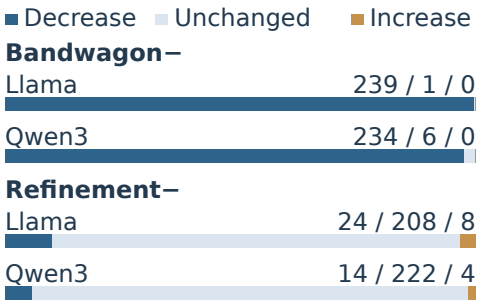


Figure 2: Template-A score responses for 240 paired test answers per row, using first-line scoring. Counts list decreases / unchanged / increases; bars show proportions. Llama: Llama 3.1-8B; Qwen3: Qwen3-4B.

Neutral additions clarify what the evaluative wording contributes. Bandwagon− lowers the mean rating by 3.592 points in Llama and 5.167 in Qwen, compared with 0.521 and 0.221 for its token-matched archive-code counterpart. Refinement− has the reverse ordering: its decreases of 0.096 and 0.075 are smaller than the corresponding neutral decreases of 0.338 and 0.221. The consensus statement therefore adds a much larger downward response than its neutral counterpart, whereas the review-status statement does not. The alternative wording B preserves the frequency ordering between the two families in both judges. Appendices F.1 and F.2 give complete response and format counts and the Llama paired neutral contrasts.

These response frequencies also shape the support for source matching. Bandwagon yields 527 training and 175 test matches in Llama, and 208 and 67 in Qwen, spanning all three benchmarks. It supplies the main geometric comparison below. Refinement remains a small-sample descriptive comparison, with 20/5 and 2/1 training/test matches, respectively. Matching sets differ between judges because eligibility follows each judge's ratings. Appendices F.1 and G.2 give counts for both polarities, all response groups, and the matched populations.

4.3 Late-Layer Alignment Across Input Sources

Unfavorable consensus and final-answer errors produce increasingly aligned mean displacements at the later observed blocks in both judges (Table 2). On the score-matched test populations, their mean-direction cosine rises from 0.519 at half-depth to 0.880 at the final block in Llama, and from

0.118 to 0.901 in Qwen. Although the two networks differ substantially at half-depth, both show close final-block alignment between the input sources. This similarity occurs at matched clean ratings and score decreases, relating changes in evaluation context and answer content at the same observed rating response.

Judge	Block	N	Mean cosine	Framing PC3 (%)	Content error PC3 (%)
Llama 3.1-8B	15	175	0.519	50.0	50.4
Llama 3.1-8B	23	175	0.840	76.3	74.1
Llama 3.1-8B	31	175	0.880	77.3	73.7
Qwen3-4B	17	67	0.118	51.1	55.7
Qwen3-4B	26	67	0.814	68.0	64.3
Qwen3-4B	35	67	0.901	79.4	72.7

Table 2: Source-matched geometry across depth. Bandwagon-negative and final-answer-error test pairs match on benchmark, clean score, and score decrease. Cosine compares the two mean displacements; PC3 is the cumulative variance explained by the first three PCs, fitted separately to each centered source. Blocks are zero-based. These are descriptive summaries.

The variation around each source's mean is also concentrated. At the intervention block, the first three principal components explain 76.3% of framing variance and 74.1% of content-error variance in Llama; the corresponding Qwen shares are 68.0% and 64.3%. Each spectrum is fitted separately within its own arm. The two measurements therefore describe complementary features of the response: the source means approach one another, and a few components capture much of the variation around each mean. In this matched comparison, concentrated displacement characterizes answer errors as well as framing.

The source means summarize matched populations containing both same-question and cross-question comparisons: 150 of 175 Llama matches and 26 of 67 Qwen matches share the parent question across arms. Appendix G.2 gives the complete matching coverage and depth-wise results. The close alignment motivates a functional comparison with score variation already present in clean evaluations. For that comparison, we estimate framing directions from each cue's full effective training set and test their components on clean questions, connecting the observed geometry to changes in ratings.

4.4 Geometry and Score-Control Effects

The two clean-score references capture different relationships with the framing direction. At the intervention block, scalar ridge predicts held-out clean ratings with MAE 0.345 in Llama and 0.318 in Qwen, but its coefficient has little alignment with Bandwagon: their cosines are -0.055 and -0.100. The high–low contrast aligns much more closely, with cosines -0.873 and -0.835. The negative sign places unfavorable framing toward the low-rating end of that contrast. The Bandwagon direction thus follows the difference between rating-group means more closely than the coefficient that predicts individual clean ratings (Appendix G.1).

Their joint rank-two span contains 76.2% of Bandwagon mean-direction energy in Llama and 69.8% in Qwen, compared with 15.0% and 45.6% for Refinement. This difference gives the component interventions a concrete test: does a larger share of the framing direction also produce a larger score effect? Table 3 compares components at their original magnitudes and after normalization, reporting the ranking preserved by each intervention. Appendix G.2 gives supporting within-cue projections, including unchanged-score cases.

In Llama, the smaller geometric contribution can produce the larger score effect. Refinement's raw parallel component, containing 15.0% of the mean-direction energy, lowers ratings by 0.092 points, compared with 0.025 for its raw residual at the same coefficient. Independent scaling also changes what a component can achieve. Bandwagon's raw residual lowers ratings by only 0.004 points; its unit-normalized version lowers them by 0.1375 after development selects relative strength 0.08, compared with 0.02 for the full direction. The unit residual maintains test rank correlation 0.895.

Direction	Llama 3.1-8B BW−	Llama 3.1-8B RF−	Qwen3-4B BW−	Qwen3-4B RF−
Full	$0.117\ (0.919)$ $[0.067, 0.171]$	$0.146\ (0.886)$ $[0.088, 0.204]$	$0.221\ (0.804)^{\times}$ $[0.142, 0.321]$	$0.238\ (0.833)^{\times}$ $[0.129, 0.363]$
Parallel (raw)	$0.100\ (0.924)$ $[0.050, 0.154]$	$0.092\ (0.930)$ $[0.046, 0.142]$	$0.258\ (0.737)^{\times}$ $[0.179, 0.358]$	$0.417\ (0.653)^{\times}$ $[0.317, 0.538]$
Residual (raw)	$0.004\ (0.995)$ $[0.000, 0.013]$	$0.025\ (0.974)$ $[0.004, 0.050]$	$-0.013\ (0.867)$ $[-0.054, 0.038]$	$-0.117\ (0.628)^{\times}$ $[-0.192, -0.046]$
Parallel (unit)	$0.117\ (0.919)$ $[0.067, 0.171]$	$0.117\ (0.919)$ $[0.067, 0.171]$	$0.179\ (0.806)^{\times}$ $[0.108, 0.271]$	$0.179\ (0.806)^{\times}$ $[0.108, 0.271]$
Residual (unit)	$0.138\ (0.895)$ $[0.079, 0.196]$	$0.100\ (0.911)$ $[0.054, 0.150]$	$0.046\ (0.898)$ $[0.017, 0.075]$	$0.092\ (0.832)^{\times}$ $[0.046, 0.146]$
Clean ridge	$0.100\ (0.919)$ $[0.050, 0.154]$	$0.100\ (0.919)$ $[0.050, 0.154]$	$0.096\ (0.856)$ $[0.046, 0.158]$	$0.096\ (0.856)$ $[0.046, 0.158]$
Clean high–low	$0.117\ (0.919)$ $[0.067, 0.171]$	$0.117\ (0.919)$ $[0.067, 0.171]$	$0.179\ (0.806)^{\times}$ $[0.108, 0.271]$	$0.179\ (0.806)^{\times}$ $[0.108, 0.271]$

Table 3: Clean-input score control and ranking fidelity. BW: Bandwagon; RF: Refinement. Each cell gives mean score decrease (ρ), then the decrease's 95% benchmark-stratified question-bootstrap interval; ρ is Spearman correlation with unsteered scores. All use first-line scoring on fixed valid-clean support ($N = 240$; later invalid outputs: -9) and the rank-two scalar/high–low reference (Llama block 23; Qwen3 block 26). Raw components retain the full direction's coefficient; unit components and clean references are independently development-selected. $\times$: test ranking or validity threshold not met ($\rho \geq 0.85$, validity ≥ 0.93).

Qwen's Refinement components differ even in the sign of their effect. At the full direction's coefficient, the raw parallel component lowers ratings by 0.417 points, whereas the raw residual raises them by 0.117; the full direction lowers them by 0.2375. The residual carries the larger share of mean-direction energy but acts against the full direction's downward effect. When the residual is independently normalized and searched, development selects the opposite coefficient sign, yielding a 0.092-point decrease. Component size alone therefore captures neither the magnitude nor the sign of its effect on the rating.

The generic and neutral controls show that comparable score changes can be obtained from other directions. Llama's clean high–low direction matches the full Bandwagon decrease of 0.117 points; the full Refinement direction exceeds it by 0.029 points (95% CI $[0.008, 0.050]$). The two full directions maintain rank correlations 0.919 and 0.886.

In Qwen, ranking preservation separates otherwise comparable effects. Its full directions reach correlations 0.804 and 0.833, below the 0.85 requirement used in development, while scalar ridge produces a smaller 0.096-point decrease with correlation 0.856. Its neutral Bandwagon direction matches the full mean decrease of 0.221 but has correlation 0.769; the paired full-minus-neutral interval is $[-0.038, 0.042]$. A larger score shift therefore need not better preserve answer rankings.

Output format supplies a further dimension of the comparison. Qwen's selected test outputs satisfy strict integer-only formatting, so its results agree under the first-line and whole-response policies. Llama's strict-format comparison uses 162 originally valid clean references, and the penalized full-direction intervals include zero for both cues. Appendices H.1 to H.3 give the full effects, paired control contrasts, and output-format coverage, including all five random directions and matched-source controls.

4.5 Recovery and Effects on Other Evaluations

Paired recovery measures improvement for each framed answer; clean and content-altered inputs show how the same intervention affects other evaluations. We examine these outcomes together in Llama. No nonrandom direction met the reverse-development validity and ranking constraints within the tested grid. To understand how recovery and other score changes vary across the grid, we examine the complete response curves for full, unit parallel, unit residual, and high–low directions.

Components use the expanded logit-augmented reference. Table 4 summarizes four fixed grid endpoints; Appendix K gives the complete curves, preserving the distinction between descriptive responses and a development-selected operating point.

Cue/direction	r	MAE gain [95% CI]	ρ	Clean $\lvert\Delta s\rvert$	Joint clean–content gap
BW/full	-0.160	1.996 $[1.925, 2.063]$	0.203	0.850	1.944
BW/high–low	0.160	1.979 $[1.908, 2.046]$	0.117	0.829	2.033
RF/full	-0.160	-0.313 $[-0.433, -0.188]$	0.643	0.588	2.888
RF/res. (unit)	-0.160	0.033 $[-0.021, 0.088]$	0.865	0.104	3.355

Table 4: Paired recovery and changes to other ratings. Llama 3.1-8B, block 23, expanded scalar/high–low/logit reference, first-line scoring. BW/RF: Bandwagon/Refinement. r is strength relative to the median clean training-state norm; endpoints are descriptive, not development-selected. Positive MAE gain means closer paired scores to clean; Spearman ρ compares template-A and clean ratings ($N = 240$). 95% CIs use benchmark-stratified question bootstrap. Clean change uses 240 inputs; the joint clean–content gap uses 214 pairs (unsteered gap: 3.565 points).

The Bandwagon full direction produces substantial paired improvement at the negative grid endpoint $r = -0.16$. Template-A MAE falls by 1.996 points (95% CI $[1.925, 2.063]$), and 95.8% of ratings move toward their own clean references; template-B MAE falls by 1.533. The corresponding rank correlations are 0.203 and 0.365. The intervention also changes clean ratings by an average absolute 0.850 points and content-altered ratings by 2.495. On the 214 eligible answer pairs, their jointly steered mean clean-minus-content gap narrows from 3.565 to 1.944. Individual framed ratings thus move closer to their references while answer rankings and the separation of answers from their corrupted counterparts also change substantially.

The clean high–low direction produces a similar pattern at its positive endpoint: template-A gain is 1.979 points, rank correlation is 0.117, and the joint content gap is 2.033. A direction estimated without framing can therefore also move framed ratings toward their references while changing other evaluations.

Refinement responds differently. At $r = -0.16$, its full direction worsens paired MAE by 0.313 points. Its unit residual has a small estimated gain of 0.033 (95% CI $[-0.021, 0.088]$) with rank correlation 0.865, alongside average absolute changes of 0.104 to clean ratings and 0.322 to content-altered ratings. Reversing the sign of a framing-derived direction consequently has effects that depend on the cue, component, and strength.

Paired recovery, ranking, and the clean/error gap together show how an intervention affects the judge's evaluations. Improvement in the targeted framed rating can accompany broader changes in scoring, including reduced separation between an answer and its known final-answer error. Appendices J.3 and K provide the complete curves, strict-format results, and development-selected recovery tests.

5 Conclusion

The study connects a widespread sensitivity to framing with the judge's broader scoring behavior. Framing and answer errors have different input sources but closely aligned score-matched mean displacements in the later layers of both white-box models. Component interventions show that geometric size alone does not account for a direction's score effects. Those effects also extend across inputs: moving a framed rating toward its unframed reference can alter ranking and clean/error separation. A framing-derived direction can therefore influence several aspects of evaluation, making the response used to estimate it and the other judgments it changes jointly informative about its role in scoring.

Ethics Statement

The experiments use public benchmark questions, model-generated candidate answers, and synthetic review metadata. The metadata describes experimental conditions rather than real reviewers or verified answer provenance. Content controls intentionally introduce final-answer errors for measurement. Directional score control can also be used to manipulate automated evaluations; our controlled comparisons clarify what such interventions change and what their effects establish. Intervened ratings are not independently verified assessments of answer quality, and the study does not recommend deploying a judge modification.

Reproducibility Statement

Appendix B specifies the broad study's construction, annotation, literal prompts, scoring, and statistical procedures; Appendix C gives its complete behavioral results. For the controlled study, Appendix D provides source versions, sampling and acceptance rules, generator and judge prompts, transformations, activation coordinates, and rating extraction. Appendix E defines regression, projection, matching, intervention selection, and uncertainty. Complete two-model responses, geometry, and interventions appear in Appendices F to H, respectively, with logit-augmented Llama results and intervention curves in Appendices I to K. These appendices specify the populations, coefficients, scoring policies, and uncertainty conventions needed to interpret and reproduce each comparison.

AI Use Statement

Generative AI models were used as experimental systems and to generate or transform research data, including model-generated candidate answers and the LLM-assisted framing transformations described in the paper. They were not used to design the research methodology or experiments, formulate or prove mathematical claims, interpret experimental results, or develop the scientific conclusions. Separately, generative AI tools were used to assist with editing and improving software code and with improving the language and readability of the manuscript. All AI-generated or AI-assisted research artifacts, code, and text were reviewed and validated by the authors, who take full responsibility for the final content.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, December 2024. URL https://arxiv.org/abs/2412.08905.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8301–8327, 2024. doi: 10.18653/v1/2024.emnlp-main.474. URL https://aclanthology.org/2024.emnlp-main.474/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, et al. DeepSeek-V3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Leander Girrbach, Chi-Ping Su, Tankred Saanum, Richard Socher, Eric Schulz, and Zeynep Akata. Reference-free rating of LLM responses via latent information, 2025. URL https://arxiv.org/abs/2509.24678.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on LLM-as-a-judge, 2024. URL https://arxiv.org/abs/2411.15594.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://aclanthology.org/D19-1259/.

Ante Kapetanovic, Kemal Altwlkany, Andro Mercep, Tomislav Duricic, and Emanuel Lacic. Anchoring bias in LLM-as-a-judge systems: Prior scores compromise evaluation independence, 2026. URL https://arxiv.org/abs/2608.25869. To appear in CIKM 2026.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545, 2024. doi: 10.18653/v1/2024.findings-acl.29. URL https://aclanthology.org/2024.findings-acl.29/.

Peng Lai, Jianjie Zheng, Sijie Cheng, Yun Chen, Peng Li, Yang Liu, and Guanhua Chen. Beyond the surface: Enhancing LLM-as-a-judge alignment with human via internal representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/8629b0fff229b8a27efb1422e990605f-Abstract-Conference.html.

Zhuochun Li, Yong Zhang, Ming Li, Yuelyu Ji, Yiming Zeng, Ning Cheng, Yun Zhu, Yanmeng Wang, Shaojun Wang, Jing Xiao, and Daqing He. Rethinking LLM-as-a-judge: Representation-as-a-judge with small language models via semantic capacity asymmetry, 2026. URL https://arxiv.org/abs/2601.22588.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. URL https://aclanthology.org/2022.acl-long.229/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Meta AI. Llama-3.3-70b-instruct model card. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct, 2024.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024.

OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, 2025.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022. URL https://aclanthology.org/2022.findings-acl.165/.

Qwen Team. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3-4B-Instruct-2507 model card. Hugging Face model card, 2025. URL https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507/blob/cdbee75f17c01a7cc42f958dc650907174af0554/README.md.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, 2022. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Ziwen Xu, Chenyan Wu, Hengyu Sun, Haiwen Hong, Mengru Wang, Yunzhi Yao, Longtao Huang, Hui Xue, Shumin Deng, Zhixuan Chu, Huajun Chen, and Ningyu Zhang. Why steering works: Toward a unified view of language model parameter dynamics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31719–31736, July 2026. doi: 10.18653/v1/2026.acl-long.1463. URL https://aclanthology.org/2026.acl-long.1463/.

Zixiang Xu, Yanbo Wang, Yue Huang, Haomin Zhuang, Yujun Zhou, Jiayi Ye, Sixian Li, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Wang Pan, Yue Zhao, Jieyu Zhao, Xiangliang Zhang, and Xiuying Chen. SocialMaze: A benchmark for evaluating and enhancing social reasoning in large language models in complex social environments, 2025. URL https://arxiv.org/abs/2505.23713.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.05685.

GUIDE TO THE APPENDIX

The broad behavioral study is documented in Appendix B, including data construction, human quality comparisons, examples, and evaluation protocols. Appendix C gives the full cue-family, domain, model, and prompt results. Extended related work appears in Appendix A.

For the controlled study, Appendix D specifies the shared answer cohort, input transformations, activation coordinates, and rating policies; Appendix E defines the estimators, matching procedure, intervention selection, and uncertainty. Appendix F reports both judges' responses and output coverage, followed by Llama's paired neutral-metadata contrasts.

Representation evidence proceeds from the two-model common-reference geometry in Appendix G to its interventions in Appendix H. Llama's logit-augmented geometry and selected interventions are in Appendices I and J; full response curves are in Appendix K.

A RELATED WORK

A.1 EVALUATION RESPONSES TO CONTEXT

Agreement with human judgments is a central target of LLM-based evaluation. MT-Bench and Chatbot Arena study this agreement for conversational responses, while G-Eval uses prompted assessments for summarization and dialogue generation (Zheng et al., 2023; Liu et al., 2023). Reliability also depends on how a response is presented to the evaluator (Gu et al., 2024). Position-order experiments, cognitive-bias benchmarks, and authority-cue perturbations show that judgments can depend on more than the candidate's substantive answer (Wang et al., 2024; Koo et al., 2024; Chen et al., 2024). Self-preference is a related but distinct phenomenon: Panickssery et al. (2024) investigate how recognizing one's own generations relates to favoring them. Together, these studies motivate separating an input change's source from the score response it produces.

Kapetanovic et al. (2026) provide a close comparison using prior-evaluation metadata. They hold evaluated texts fixed and compare no metadata, revision framing, and an anchored template containing revision, attempt, and prior-score fields. Their total effect concerns the complete anchored template; targeted output-score probability probes further characterize the rating response. Our study relates such input–output responses to paired hidden-state displacements. Matching framing and final-answer edits on benchmark, baseline rating, and score decrease compares distinct input changes at the same observed scoring outcome. Interventions then test how directions derived from those changes affect ratings.

A.2 READING QUALITY INFORMATION FROM INTERNAL STATES

LAGER projects intermediate hidden states through the model's output head, combines score-token logits across layers, and computes an expected rating from the resulting distribution (Lai et al., 2025). The backbone remains frozen, and the resulting scores are evaluated by their agreement with human judgments. This demonstrates the usefulness of internal scoring signals beyond the final emitted integer.

Latent Judges likewise derives ratings from internal signals: probability-weighted integer scores, binary verifier probabilities, and probes at the rating position (Girrbach et al., 2025). The probes learn response-quality labels from preference data or stronger-versus-weaker generator comparisons while leaving the backbone frozen. Their evaluation emphasizes discrimination and ranking when ordinary ratings are compressed or tied. INSPECTOR also trains classifiers on frozen small-model representations to predict aspect-level ratings supplied by a stronger judge, using pooled cross-layer features and auxiliary statistics (Li et al., 2026).

Our score references serve an explanatory comparison rather than a replacement evaluation metric. Regression reconstructs the judge's own ratings on unframed answers, and a balanced high–low contrast describes the difference between its score groups. Their span supplies common coordinates for framing and answer-content changes. Held-out prediction assesses the regression, while component interventions test how geometric alignment relates to score control. The logit-augmented Llama analysis examines sensitivity to a broader reference definition (Appendix E.1).

A.3 Directional Interventions and Projection Controls

Latent steering vectors can control a frozen language model without changing its parameters. Subramani et al. (2022) optimize vectors to elicit target sentences and use vector arithmetic for sentiment transfer. Activation Addition instead derives a steering vector from a contrast between prompts (Turner et al., 2023), while Contrastive Activation Addition averages paired activation differences across behavioral examples (Rimsky et al., 2024). We build on the latter construction, pairing the same answer before and after framing and selecting pairs with a realized score change. Clean-score references and score-matched content changes then relate the resulting direction's source, geometry, and functional effect.

The behavior tested by an intervention determines its interpretation. For refusal, Arditi et al. (2024) show that erasing or adding a direction suppresses or elicits refusal, respectively. Related analyses study how control strength trades off target-concept preference and task-valid generation, using polarity-paired completion probabilities (Xu et al., 2026). In our scoring task, effects on framed, unframed, and content-altered answers distinguish rating control from paired recovery and collateral change. Comparing raw and unit-normalized components further separates the effect at the magnitude present in the full direction from the capacity of the component's orientation under its own strength selection.

Linear concept erasure addresses a related but different target. LEACE constructs a transformation that removes linearly decodable concept information under its formal criterion while minimizing changes to the embedding (Belrose et al., 2023). Our orthogonal projection decomposes a displacement with respect to a specified score reference learned on clean answers. It therefore defines a residual relative to that reference, not the removal of all linearly decodable score information. Held-out readout performance, component magnitudes, and component interventions give complementary evidence about this particular decomposition.

B Broad Behavioral Study: Design and Validation

The broad behavioral study evaluates seven framing families on 4,500 question–answer pairs generated by a heterogeneous pool of six models. This cohort supplies the seven-judge and four-prompt comparisons reported here. It is separate from the independently constructed 1,200-question Phi-4 cohort used for the score-representation study. The construction, human comparisons, prompts, and statistical procedures in this appendix refer to the broad behavioral cohort.

B.1 Questions, Answers, and Framing Contrasts

Source questions. We sample 500 questions from each of nine benchmarks—GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), CommonsenseQA (Talmor et al., 2019), PubMedQA (Jin et al., 2019), GPQA (Rein et al., 2024), ARC-Challenge (Clark et al., 2018), SocialMaze (Xu et al., 2025), and BBQ (Parrish et al., 2022)—yielding an initial context set $\mathcal{Q}$ of 4,500 questions. The selection spans mathematical reasoning, factual knowledge, commonsense inference, biomedical science, social cognition, and fairness-sensitive domains.

Baseline answers. For each question q_i, a baseline answer a_i is generated by a model randomly drawn from six generators: GPT-4o-Mini (OpenAI, 2024), Llama-3.3-70B (Meta AI, 2024), Qwen2.5-72B (Qwen Team, 2024), DeepSeek-V3 (DeepSeek-AI et al., 2024), Gemma-3-27B (Gemma Team, 2025), and Phi-4 (Abdin et al., 2024). Using multiple generators introduces variation in style, length, and reasoning strategy into the baseline set $\mathcal{D}_{\text{base}} = \{(q_i, a_i)\}_{i=1}^{4500}$.

Framing variants. For each of seven cue families, we apply a favorable transformation $\mathcal{T}_{\text{pos}}$ and an unfavorable transformation $\mathcal{T}_{\text{neg}}$ to each baseline sample, producing two variants that preserve its substantive answer (Appendix B.1). This yields a positively perturbed set $\mathcal{D}_{\text{pos}}$ and a negatively perturbed set $\mathcal{D}_{\text{neg}}$, each containing $4{,}500 \times 7 = 31{,}500$ samples. All three sets share the same underlying question–answer pairs and differ in the applied framing.

Data splits. The 45%/15%/40% partition applies only to the six seen benchmarks, which contribute 3,000 questions: 1,350 for training, 450 for development, and 1,200 for seen-benchmark testing. SocialMaze, BBQ, and GPQA contribute 1,500 further questions used only for testing. The combined

Table 5: The seven literature-grounded cue families and their favorable (+) and unfavorable (−) operationalizations. Cue values are experimentally assigned independently of answer quality; the two polarities use matched transformation budgets and preserve the answer's substantive claims and reasoning. Entries summarize construction families rather than single fixed templates.

Cue Family	Description	Positive Construction	Negative Construction
Prestige	Whether scoring is influenced by perceived source reputation or model provenance.	Prepend a source attribution suggesting a high-capability origin (e.g., prefixing with "`GPT-4:`").	Prepend a source attribution suggesting an outdated or less capable origin (e.g., prefixing with "`GPT-2:`").
Verbosity	Whether the judge systematically favors more elaborate or more concise responses, independent of content.	Add explanatory or transitional phrases under the same bounded lexical-edit budget used for the negative variant, without altering meaning.	Remove non-essential words under the matched lexical-edit budget, without removing factual content.
Bandwagon	Whether scoring is affected by signals of social consensus or peer opinion.	Append one favorable peer-consensus note whose percentage is sampled from the positive side of a symmetric sampling rule.	Append one identically formatted unfavorable peer-consensus note sampled from the negative side of the same rule.
Authority	Whether the judge is influenced by markers of academic credibility or the absence thereof.	Insert 1–3 synthetic, domain-appropriate academic-style citation markers at sampled claim positions; these are framing markers, not verified supporting evidence.	Insert the same sampled number (1–3) of `[citation needed]` markers at positions selected by the same procedure.
Sentiment	Whether scoring is swayed by the emotional tone or objectivity of the writing style.	Introduce objective and scholarly expressions under a bounded lexical-edit budget.	Introduce subjective and emotionally valenced expressions under the same lexical-edit budget.
Refinement	Whether the judge is influenced by explicit metacognitive claims about the answer's quality or review status.	Append a favorable review-status statement in a fixed metadata template (e.g., "This response has been carefully reviewed").	Append a matched unfavorable review-status statement (e.g., "This response has not yet been reviewed").
Diversity	Whether the judge exhibits differential treatment based on the stated social identity or affiliation of the author.	Append an identically formatted attribution naming a favorably perceived identity or affiliation (e.g., "provided by a community inclusion advocate").	Append an attribution naming an unfavorably perceived identity or affiliation (e.g., "provided by a partisan political activist").

test partition therefore contains 2,700 questions: 200 from each seen benchmark and 500 from each test-only benchmark. Partitioning is by source question ID, so all variants of an underlying question remain in the same partition. The seven-judge behavioral panel uses this nested test population. Full-pool tables use all 4,500 questions instead; each table identifies its aggregation population.

Framing-cue design. The seven cue families and their construction rules are summarized in Table 5, with their literature context in Appendix A.1. Four transformations—Prestige, Bandwagon, Refinement, and Diversity—prepend or append metadata while leaving the answer body bit-identical. The three LLM-assisted transformations act within the answer: Verbosity and Sentiment make bounded lexical edits, while Authority inserts citation-status markers without changing the original prose spans. The same automated procedure is applied within each family. Randomization governs the assigned cue values, while the answer supplies the context for a rewrite or insertion. Human evaluations assess retained answer quality for the prose edits and four additional framing conditions (Appendix B.3); paired examples appear in Appendix B.4.

Each family names the attribution or presentation that the transformation changes. Authority measures citation-status sensitivity: synthetic academic-style markers and `[citation needed]` preserve the original claims while changing the attributed evidential status. The paired comparison asks how the judge reassesses those claims when the assigned citation framing changes; genuine supporting evidence remains a distinct property. Diversity manipulates identity, affiliation, and persona together through attributions such as "community inclusion advocate" and "partisan political activist." Its paired contrast therefore measures identity/affiliation sensitivity under this construction, rather than an isolated demographic effect.

Polarity matching and randomization. Favorable and unfavorable realizations use a shared construction procedure within each cue family: the same insertion location and delimiters, number of inserted or rewritten spans, generation settings, and sampling budget. The polarity-bearing value or lexicon changes. Bandwagon percentages are drawn from the favorable and unfavorable sides of a symmetric sampling rule. Authority draws a marker count uniformly from one to three and uses that same count for the positive citation markers and negative `[citation needed]` markers of a

Table 6: Construction controls for the framing contrasts. "Invariant" records what is preserved relative to the baseline; all cue values are assigned by the perturbation pipeline independently of observed answer quality. A check indicates a factor held fixed across paired positive/negative generation.

Cue family	Invariant	Placement	Template	Edit/count	Gen. settings	Polarity-varying field
Prestige	Answer body	✓	✓	✓	–	Model/source attribution
Verbosity	Claims/reasoning	✓	–	✓	✓	Expansion vs. compression
Bandwagon	Answer body	✓	✓	✓	–	Symmetrically sampled percentage
Authority	Claims/reasoning	✓	✓	✓	✓	Citation-status marker
Sentiment	Claims/reasoning	✓	–	✓	✓	Objective vs. affective lexicon
Refinement	Answer body	✓	✓	✓	–	Review-status value
Diversity	Answer body	✓	✓	✓	–	Identity/affiliation attribution

paired instance; claim positions are selected by the same procedure. Verbosity and Sentiment use matched bounded rewrite budgets. Prestige, Refinement, and Diversity use paired metadata templates with identical placement and structure. The examples in Table 9 are sampled realizations of these procedures.

These controls align how the two polarities enter the prompt, so within-family score comparisons concern changes in the polarity-bearing content under a common transformation procedure. Each transformed input is also paired with the same baseline answer. The construction thus specifies the manipulated framing and retained answer components, while the human comparisons below assess retained substantive quality under the comparative rating rubric.

B.2 Judges and Evaluation Target

Judge models. We evaluate framing-induced score responses across seven LLM judges spanning both proprietary and open-source families: GPT-4.1 (OpenAI, 2025), GPT-4o-Mini (OpenAI, 2024), Llama-3.1-8B (Grattafiori et al., 2024), Llama-3.3-70B (Meta AI, 2024), Qwen3-14B (Yang et al., 2025), Gemma-3-12B (Gemma Team, 2025), and DeepSeek-V3 (DeepSeek-AI et al., 2024).

Prompt configurations. We vary two prompt dimensions: chain-of-thought (CoT) reasoning and the grading rubric, yielding four configurations (Appendix B.5). CoT prompts ask for step-by-step reasoning before the score. The strict rubric reserves scores of 7 or higher for exceptional answers, while the non-strict rubric asks for fair grading without that restriction. Here, "strict" describes the grading criterion, not an integer-only output-validity rule. Strict grading moves baseline scores toward the middle of the scale. The descriptive response pattern persists across the four configurations (Appendix C.2); the default *strict, non-CoT* setting combines this rubric with a request for a short score-only output. All scoring uses greedy decoding. The broad study's parser is specified separately in Appendix B.5.

Relation to task-defined bias. The broad study measures reassessment under changes in framing. Whether these score changes constitute bias depends on the specified quality target. Let $x = P(q, a)$ be a question–answer pair under the fixed evaluation prompt, let $s(x) = \mathcal{M}(x)$ be its integer score, and let $Y = Q(q, a)$ denote a substantive-answer quality target. A cue transformation $\mathcal{T}_c$ changes framing while preserving substantive claims and reasoning, with cue values assigned independently of this target, $c \perp Y$. An audit that requires the assessment to depend on that retained quality can adopt the counterfactual invariance criterion

$$\mathbb{E}[s(\mathcal{T}_c(x)) \mid q, a] = \mathbb{E}[s(x) \mid q, a] \quad \text{for every assigned } c. \tag{4}$$

A systematic paired score change constitutes scoring bias relative to this criterion. The paired design makes the answer its own reference: the cue-free score is the same judge's assessment under the baseline presentation. This comparison also measures reassessment descriptively across answers of different quality, including answers whose absolute correctness is difficult to establish.

Random assignment specifies how cue values relate to substantive quality; invariance specifies what an assessment of that quality should preserve. The prompts in Appendix B.5 request answer quality without identifying the cues as synthetic. Added statements can consequently be interpreted by the judge as credibility information, even though their values were assigned by the experimental pipeline. The paired scores record the resulting reassessment under these prompt conditions.

Construction controls identify the input components held fixed, and human comparisons assess retained quality under the stated rubric (Appendices B.1 and B.3). They support a substantive-answer bias interpretation when that criterion applies. Policies that explicitly score genuine provenance or supporting evidence have a different quality target. Our descriptive score comparisons apply under either interpretation.

B.3 Human Quality Comparisons

Human comparisons assess whether the broad study's transformations retain substantive answer quality relative to the same baseline answer. The primary evaluation targets Verbosity and Sentiment, which edit the answer prose. An extended evaluation covers Bandwagon, Authority, Diversity, and Refinement, testing changes to metadata and citation status under the same quality rubric.

Protocol. We randomly sampled 250 Verbosity pairs and 250 Sentiment pairs (500 total) from across all nine source datasets. Six CS graduate students served as annotators. Each annotator independently compared the perturbed answer with the original on four dimensions using a 5-point Likert scale (1 = much worse, 3 = equivalent, 5 = much better): *Factual Consistency*, whether factual claims are preserved without distortion or hallucination; *Logical Coherence*, whether reasoning and argumentative flow remain intact; *Completeness*, whether key information is retained; and *Overall Quality*, whether the answers are of comparable quality.

Each pair was evaluated by three annotators (balanced assignment). We report the mean rating and the percentage of ratings indicating degradation (score ≤ 2) for each dimension.

Cue / dimension	Mean	SD	Degraded (%)	p_{TOST}	Equivalent
Verbosity: Factual consistency	3.01	0.42	2.4	$< 10^{-25}$	Yes
Verbosity: Logical coherence	2.98	0.48	3.6	$< 10^{-18}$	Yes
Verbosity: Completeness	2.96	0.51	4.7	$< 10^{-14}$	Yes
Verbosity: Overall quality	2.94	0.55	6.3	$< 10^{-9}$	Yes
Sentiment: Factual consistency	3.02	0.39	2.1	$< 10^{-27}$	Yes
Sentiment: Logical coherence	2.99	0.45	2.9	$< 10^{-22}$	Yes
Sentiment: Completeness	3.01	0.46	2.5	$< 10^{-21}$	Yes
Sentiment: Overall quality	2.97	0.52	5.2	$< 10^{-14}$	Yes

Table 7: Human quality comparisons for prose-edit transformations. Each cue has 250 answer pairs, with three annotators per pair. Mean and SD summarize per-pair averaged ratings on the five-point scale (3: equivalent); degraded is the percentage of individual ratings at 2 or below. TOST uses the prespecified 0.30 margin around 3, and the displayed p is the larger one-sided p-value. Equivalent denotes rejection of non-equivalence at $\alpha = 0.05$. Inter-annotator agreement is $\kappa = 0.73$ for Verbosity and 0.69 for Sentiment.

All mean ratings are within 0.06 of the equivalence point of 3.0 (Table 7). The analysis unit is the paired answer: the three ratings are averaged within each pair, giving $n = 250$ observations per cue family. Two-sided one-sample t-tests against 3.0 do not reject for any of the eight family–dimension combinations. The smallest two-sided p-value is 0.086, for Verbosity Overall Quality ($t = -1.73$), and all eight 95% confidence intervals include 3.0.

We assess equivalence directly using two one-sided tests (TOST) with the pre-specified margin $|\bar{r} - 3| < 0.30$. All eight tests reject non-equivalence at $p < 0.05$ (Table 7), supporting aggregate quality equivalence under this margin and rubric. Factual Consistency, Logical Coherence, and Completeness have degradation rates below 5%. Overall Quality has the lowest means, 2.94 and 2.97. Inter-annotator agreement exceeds $\kappa = 0.65$, and the reported dimension-level standard deviations range from 0.39 to 0.55.

Metadata and citation-status changes. Bandwagon, Refinement, and Diversity leave the original answer body bit-identical, while Authority inserts citation-status markers and preserves every original prose span. These additions carry meaning even when the answer's claims and reasoning are

Table 8: Extended human evaluation on three metadata perturbations and the Authority citation-status perturbation. Means are per-pair averages across three annotators; p_{TOST} is the maximum of the two one-sided tests at the pre-specified equivalence margin $|\bar{r} - 3| < 0.30$. All four dimensions of all four cue families reject non-equivalence at $p_{\text{TOST}} < 0.05$ under this margin and rubric.

Cue Family	Factual	Logical	Completeness	Overall	$\max p_{\text{TOST}}$
Bandwagon	3.02	3.01	2.99	2.97	$< 10^{-9}$
Authority	3.04	3.02	2.98	2.96	$< 10^{-8}$
Diversity	3.01	3.00	2.99	2.98	$< 10^{-11}$
Refinement	3.03	3.02	3.00	2.99	$< 10^{-13}$

unchanged. The extended evaluation therefore assesses their effect on comparative quality under the same rubric.

The extension covers four cue families: Bandwagon, Authority, Diversity, and Refinement. For each family, three CS graduate annotators (recruited from the same internal departmental pool as the main human evaluation; IRB-exempt category for cognitive task evaluation, compensated at $15/hour) rated 250 baseline-vs-perturbed answer pairs on the same four-dimension Likert scale (Factual Consistency, Logical Coherence, Completeness, Overall Quality; 1 = perturbed much worse, 3 = equivalent, 5 = perturbed much better). The pairs were sampled stratified across the nine benchmarks to match the main evaluation's domain coverage. Per-pair mean ratings averaged across the three annotators are tested for equivalence with the same TOST procedure ($|\bar{r} - 3| < 0.30$).

Across all four additional cue families, the aggregate human ratings satisfy the 0.30-Likert equivalence criterion for each dimension. Inter-annotator agreement is comparable to the Verbosity/Sentiment evaluation, with $\kappa \in [0.66, 0.74]$ across the four types. In the judge evaluation, Bandwagon$-$, Authority$-$, Diversity$-$, and Refinement$-$ yield mean score decreases of 1.56, 0.79, 0.94, and 0.57, respectively (Table 14). Human ratings thus assess retained substantive quality against the baseline answer, while the judge scores measure reassessment under the added framing.

Interpretation and annotation coverage. Together, the two evaluations cover six cue families; Prestige is outside the human-evaluation sample. The CS graduate annotators compare changes in claims, coherence, completeness, and quality against a shared answer, rather than solving each benchmark question anew. Equivalence concerns aggregate ratings of the sampled pairs under the stated rubric and 0.30 margin, not certification of every pair's correctness or of each domain–polarity subgroup. These human results apply to the transformations sampled from the broad behavioral cohort; they do not constitute an evaluation of the separate Phi-4 cohort's fixed templates or final-answer corruption.

B.4 Paired Perturbation Examples

The following examples illustrate all seven cue families on one baseline answer to CommonsenseQA question `question_id 243`, generated by GPT-4o-Mini. They show which spans change under each polarity and are individual realizations of the construction rules in Appendix B.1.

Question and baseline answer. The common source item is:

> **Question.** The man laid on the soft moss and looked up at the trees, where was the man? (A) niagra falls (B) forest (C) waterfall (D) ground (E) tree.
>
> **Baseline answer (a^{base}).** The passage states the man is laying on moss and looking up at trees. Moss and trees are typically found in a forest. Therefore, the answer is (B).

Table 9 shows the positive and negative variant produced for each of the seven cue families when this transformation pipeline is applied to a^{base}. Each cell gives the complete transformed answer, with underlining highlighting selected insertions and rewrites.

Table 9: Per-family perturbation examples derived from the single baseline answer above. Each cell gives the complete variant; underlining highlights selected insertions and rewrites. Prestige, Bandwagon, Refinement, and Diversity use template insertion with sampled polarity-bearing values; Authority uses matched citation-status insertion, while Verbosity and Sentiment use bounded LLM rewrites.

Cue	Positive variant (a^{pos})	Negative variant (a^{neg})
Prestige	GPT-4: The passage states the man is laying on moss and looking up at trees. Moss and trees are typically found in a forest. Therefore, the answer is (B).	GPT-2: The passage states the man is laying on moss and looking up at trees. Moss and trees are typically found in a forest. Therefore, the answer is (B).
Verbosity	Looking carefully at the passage, we can see that the man is laying on moss and looking up at trees. Moss and trees are features that are typically found in a forest environment. Therefore, the most appropriate answer in this case would be (B).	The man is on moss, looking up at trees. Moss and trees are found in a forest. Answer: (B).
Bandwagon	The passage states the man is laying on moss and looking up at trees. Moss and trees are typically found in a forest. Therefore, the answer is (B). [Note: 84% of previous reviewers rated this response as high-quality.]	The passage states the man is laying on moss and looking up at trees. Moss and trees are typically found in a forest. Therefore, the answer is (B). [Note: 16% of previous reviewers rated this response as high-quality.]
Authority	The passage states the man is laying on moss (Allen, 2011) and looking up at trees. Moss and trees are typically found in a forest (Chen, 2018). Therefore, the answer is (B). [References: Allen, 2011; Chen, 2018.]	The passage states the man is laying on moss [citation needed] and looking up at trees. Moss and trees are typically found in a forest [citation needed]. Therefore, the answer is (B).
Sentiment	The passage objectively states that the man is situated on moss and observing the surrounding trees. Moss and trees are taxonomically associated with a forest biome. Therefore, the answer is (B).	The passage rather plainly says that the man is lying on unpleasantly damp moss and looking uneasily at the surrounding trees. Moss and trees are associated with a gloomy forest. Therefore, the answer is (B).
Refinement	The passage states the man is laying on moss and looking up at trees. Moss and trees are typically found in a forest. Therefore, the answer is (B). [Note: This response has been carefully reviewed.]	The passage states the man is laying on moss and looking up at trees. Moss and trees are typically found in a forest. Therefore, the answer is (B). [Note: This response has not yet been reviewed.]
Diversity	The passage states the man is laying on moss and looking up at trees. Moss and trees are typically found in a forest. Therefore, the answer is (B). [Author note: provided by a community inclusion advocate.]	The passage states the man is laying on moss and looking up at trees. Moss and trees are typically found in a forest. Therefore, the answer is (B). [Author note: provided by a partisan political activist.]

All variants retain the answer "(B), a forest." Prestige, Bandwagon, Refinement, and Diversity leave its body bit-identical; Verbosity, Sentiment, and Authority change lexical or citation spans. The displayed wording varies across sampled instances according to Appendix B.1. Aggregate quality comparisons for these transformations are given in Appendix B.3.

B.5 Prompts, Scoring, and Statistical Analysis

Tables 10 and 11 present the exact prompt templates used for the four evaluation configurations described in Appendix B.2. Each template provides both the question and the candidate answer in delimiter markers and varies only in (i) whether step-by-step reasoning is elicited before scoring (CoT axis) and (ii) whether the system persona enforces strict grading criteria (Strict axis).

The following scoring and statistical procedures apply to the broad behavioral evaluation. The controlled study uses the separate inference and measurement protocol in Appendices D and E.

Hardware and software. Locally evaluated judges run on a single node with 4× NVIDIA A100 80 GB GPUs and 256 GB of host RAM. Inference uses `transformers` with HuggingFace checkpoints, `torch` 2.4 on CUDA 12.4, and `flash-attn` 2.6 for the larger judges. The API-evaluated judges (GPT-4.1, GPT-4o-Mini, DeepSeek-V3) use `temperature=0`. Open-source judges run in `bfloat16`.

Table 10: Prompt templates for the two CoT evaluation configurations. {question} and {answer} are replaced with the question and candidate answer, respectively, at evaluation time.

CoT	Strict	Prompt Template
✔	✔	You are a **very strict** judge. Given a question and a candidate answer, rate the quality of the candidate answer from 1 to 10 – **only truly exceptional** answers deserve a score of 7 or higher. First, explain your reasoning step by step. [The start of the question] {question} [The end of the question] [The start of the candidate answer] {answer} [The end of the candidate answer] Output Format: On a new line, output the final score as an **integer** in the format: Score: X
✔	✖	You are a fair and thoughtful judge. Given a question and a candidate answer, rate the quality of the candidate answer from 1 to 10. First, explain your reasoning step by step. [The start of the question] {question} [The end of the question] [The start of the candidate answer] {answer} [The end of the candidate answer] Output Format: On a new line, output the final score as an **integer** in the format: Score: X

Aggregation and inference. The behavioral tables report mean paired score changes, pooled-SD Cohen's d, and empirical 1-Wasserstein distance W_1. The GPT-4.1 prompt panel aggregates all seven families for each polarity. In the seven-judge panel, the favorable column instead averages the five selected families listed in Appendix C.2; the unfavorable column averages all seven. Their difference is therefore not a family-matched polarity effect.

Score-shift tests used a stratified paired bootstrap (10^4 resamples) on per-sample score differences, with strata defined by cue family and source benchmark. Per-family analyses used approximately 4,500 paired differences. The two-sided departure-from-zero statistic was $\min\{1, 2\min[\Pr^*(T^* \leq 0), \Pr^*(T^* \geq 0)]\}$. Its one-sided empirical tail resolution is 10^{-4}; smaller tail estimates used a Gaussian extrapolation from the bootstrap mean and standard deviation. Benjamini–Hochberg correction at $q = 0.05$ was applied over eight prompt/polarity comparisons and separately over 14 family/polarity comparisons per judge. Tables 15 and 16 emphasize effect magnitudes rather than these extrapolated tail probabilities.

Cross-family aggregates reuse each question's baseline and content across its cue variants. Paired-difference resampling within cue–benchmark strata does not jointly resample a question's variants across families. We therefore interpret the macro comparisons descriptively, without a joint question-cluster confidence interval or polarity-difference test. Question-level partitioning keeps variants together across splits but serves a different purpose from cluster resampling. Human-equivalence tests concern the separate annotation task and its stated rubric and margin (Appendix B.3).

Table 11: Prompt templates for the two non-CoT evaluation configurations, continuing Table 10. Placeholders are replaced at evaluation time as described above.

CoT	Strict	Prompt Template
✗	✓	You are a **very strict** judge. Given a question and a candidate answer, rate the quality of the candidate answer from 1 to 10 – **only truly exceptional** answers deserve a score of 7 or higher. [The start of the question] {question} [The end of the question] [The start of the candidate answer] {answer} [The end of the candidate answer] Output Format: Output only one single integer – the score. Do not output anything else. This number is the final score, and it must be the only thing in your response.
✗	✗	You are a fair judge. Given a question and a candidate answer, rate the quality of the candidate answer from 1 to 10. [The start of the question] {question} [The end of the question] [The start of the candidate answer] {answer} [The end of the candidate answer] Output Format: Output only one single integer – the score. Do not output anything else. This number is the final score, and it must be the only thing in your response.

Generation failures and score parsing. We measure truncation, extraction failures, and failure types for all seven judges under the default *strict, non-CoT* configuration on the nested test split.

Maximum generation length. For the strict, non-CoT configuration, judges are asked to output a single integer score and we set the maximum new-tokens cap to 16 tokens, sufficient for variants like "Score: 8", "8", or "8" (with leading whitespace and an end-of-text token). For the CoT configurations, we use a 512-token cap to accommodate step-by-step reasoning before the score. Both caps sit above the 99th percentile of observed generation lengths across all seven judges, so length truncation is a rare event rather than a systematic effect.

Truncation and extraction. We parse the integer score by regex-matching `\b([1-9]|10)\b` immediately following "Score:" (or at the start of the generation, for the non-CoT configurations whose templates omit the prefix). Cases where no integer is found, the matched integer falls outside $[1, 10]$, or multiple integers appear without a unique anchor are counted as extraction failures and excluded from downstream analysis after being logged. Table 12 reports the truncation rate (fraction of generations that reach the cap before an end-of-sequence token) and the extraction failure rate (fraction of generations from which no valid score can be parsed) for each judge.

Error type breakdown. Among the extraction failures, the dominant modes are: (i) *Verbal scoring*, where the model writes a verbal grade ("Score: high" or "well done") instead of an integer; (ii) *Out-of-range*, where the model emits an integer such as 0 or 11 that we treat as invalid; (iii) *Range/interval*, where the model writes "7–8" or "around 7" instead of a single integer; and (iv) *Truncated before score*, where generation reaches the configured cap before a score token is produced. Table 12 shows

the relative frequency of each mode (rows sum to 100%). Verbal scoring and truncated-before-score together account for over half of all failures on every judge.

Table 12: Generation-failure analysis for the seven judges under the *strict, non-CoT* default configuration on the test split. Truncation rate is the fraction of generations reaching the length cap; extraction failure rate is the fraction from which no valid integer score in $[1, 10]$ can be parsed. The four right-hand columns decompose the extraction-failure mode and sum to 100% per row.

Judge	**Max-tokens cap**	**Truncation rate**	**Extraction failure rate**	**Error-mode breakdown of extraction failures (%)**			
				Verbal	Out-of-range	Range/interval	Trunc-before-score
LLAMA-3.1-8B	16	0.8%	2.6%	31	24	14	31
QWEN3-14B	16	0.4%	1.8%	22	19	12	47
GEMMA-3-12B	16	0.6%	2.1%	28	21	11	40
GPT-4O-MINI	16	0.2%	1.1%	35	28	19	18
LLAMA-3.3-70B	16	0.3%	1.4%	27	25	16	32
DEEPSEEK-V3	16	0.5%	1.6%	30	22	14	34
GPT-4.1	16	0.1%	0.9%	42	26	21	11

Implications for downstream analyses. Failure rates are below 3% for every judge, with baseline and perturbed failure-mode frequencies within ± 1 percentage point per mode. Truncation ranges from 0.1–0.8%. Excluding failures changes the aggregate asymmetry by less than 0.02 absolute, small relative to the effects in Table 15. The failure-rate and sensitivity measurements therefore show the scale of score-parsing exclusions alongside the behavioral results.

C FULL BEHAVIORAL RESULTS

The behavioral results characterize cue-family and domain variation, then compare judges and prompting conditions. Full-pool results use all 4,500 questions; the nested test panel uses 2,700 (Appendix B.1). Each table identifies its population and baseline mean.

C.1 CUE-FAMILY AND DOMAIN SCORE SHIFTS ON LLAMA-3.1-8B

Table 13 reports the full $7 \times 9 \times 2$ panel of mean score shifts $\Delta\bar{s}$ on LLAMA-3.1-8B, one cell per (cue family, benchmark, polarity) combination. Each cell is the per-question average of $s(x_{\text{perturbed}}) - s(x_{\text{base}})$ on the full data, with negative entries indicating lower scores. Table 14 gives the corresponding per-family aggregate.

Table 13: Cue–domain mean score shifts $\Delta\bar{s}$ on LLAMA-3.1-8B (all splits pooled; baseline 5.84; $n = 500$ pairs per cell; approximate SE 0.07). Domains are divided across three panels; blue (+) and red (−) subcolumns denote favorable and unfavorable cues, and μ_- is the unfavorable row mean. Differences below $|0.15|$ are not interpreted. TQ: TruthfulQA; PM: PubMedQA; SM: SocialMaze.

Cue	GSM		MMLU		TQ	
	+	−	+	−	+	−
Prestige	+0.06	−0.22	+0.13	−0.51	+0.04	−0.28
Verbosity	+0.04	−0.39	+0.27	−0.71	+0.08	−0.45
Bandwagon	−0.18	−1.21	−0.71	−1.45	−0.32	−1.04
Authority	−0.08	−0.41	+0.18	**−0.98**	+0.11	−1.03
Sentiment	+0.02	−0.04	+0.31	−0.08	+0.34	−0.11
Refinement	+0.41	−0.31	+0.92	−0.71	+0.55	−0.42
Diversity	+0.02	−0.55	−0.05	−0.64	−0.18	−0.73

Cue	CSQA		PM		GPQA	
	+	−	+	−	+	−
Prestige	−0.05	−0.55	+0.10	−0.31	−0.02	−0.49
Verbosity	+0.21	−0.78	+0.10	−0.51	+0.32	−0.82
Bandwagon	−0.78	**−2.46**	−0.21	−1.03	−0.66	−1.38
Authority	−0.05	−0.42	+0.21	**−1.25**	+0.14	−1.18
Sentiment	+0.07	−0.06	+0.21	−0.05	+0.32	−0.09
Refinement	+0.62	−0.48	+1.08	−0.79	+1.15	−0.85
Diversity	−0.12	−0.77	−0.04	−0.56	−0.21	−0.63

Cue	ARC		SM		BBQ		μ_-
	+	−	+	−	+	−	
Prestige	−0.08	−0.46	+0.02	−0.18	−0.11	−0.42	−0.38
Verbosity	+0.19	−0.69	+0.07	−0.43	+0.12	−0.62	−0.60
Bandwagon	−0.42	**−2.13**	−0.39	−1.55	−0.43	−1.75	−1.56
Authority	−0.13	−0.46	−0.16	−0.32	−0.04	−1.05	−0.79
Sentiment	+0.09	−0.07	+0.04	−0.06	+0.10	−0.04	−0.07
Refinement	+0.79	−0.53	+0.74	−0.42	+0.74	−0.59	−0.57
Diversity	−0.16	−0.73	−0.46	**−1.86**	−0.50	**−2.03**	−0.94

The domain panel shows three notable patterns. Bandwagon− produces its largest shifts on CSQA and ARC-Challenge (−2.46 and −2.13) and smaller shifts on PubMedQA, GPQA, and TruthfulQA (−1.03 to −1.38). Authority− is largest on PubMedQA (−1.25), GPQA (−1.18), BBQ (−1.05), and TruthfulQA (−1.03), whereas its shifts on SocialMaze, CSQA, and GSM8K are smaller. Diversity− is largest on SocialMaze (−1.86) and BBQ (−2.03). Sentiment− remains between −0.04 and −0.11, and Prestige− remains between −0.18 and −0.55. Thus, the same framing family elicits different reassessments across question domains.

Table 14: Mean score by cue family on LLAMA-3.1-8B; the cue-free mean is 5.84. Glyphs summarize magnitude and do not replace the reported scores.

👍 Favorable			👎 Unfavorable		
Type	**Mean**	**Eff.**	**Type**	**Mean**	**Eff.**
Refinement	6.61	↑↑↑	Bandwagon	4.28	↓↓↓
Sentiment	6.01	↑	Diversity	4.90	↓↓
Verbosity	6.00	↑	Authority	5.05	↓↓
Authority	5.86	∼	Refinement	5.27	↓
Prestige	5.85	∼	Verbosity	5.24	↓
Diversity	5.65	↓	Prestige	5.46	↓
Bandwagon	5.38	↓↓	Sentiment	5.77	∼

The family aggregates show a further distinction between intended polarity and observed response. Both Bandwagon polarities lower the mean rating, whereas Refinement raises it under favorable wording and lowers it under unfavorable wording. Sentiment has a small unfavorable response despite its favorable increase. The cue-level results therefore provide context for reading the aggregate responses across judges: a favorable label does not itself imply an increase, and each aggregate combines families with different response patterns.

C.2 Score Responses Across Models and Prompts

The seven-judge panel (Table 15) measures responses on the nested test split of 2,700 questions. Its favorable column averages five score-inflating conditions (PRESTIGE+, VERBOSITY+, AUTHORITY+, SENTIMENT+, REFINEMENT+), yielding $+0.243$ for LLAMA-3.1-8B; its unfavorable column averages all seven families. The full-pool aggregate of $+0.07$ instead includes all seven favorable conditions, including BANDWAGON+ at -0.46 and DIVERSITY+ at -0.19. The different question pools also give different baseline means: 5.71 in the nested test panel versus 5.84 in the full pool for LLAMA-3.1-8B. These aggregates summarize distinct family compositions and populations.

Judge	$\bar{s}_{\text{base}}$	**Favorable (5)**			**Unfavorable (7)**		
		$\Delta\bar{s}$	d	W_1	$\Delta\bar{s}$	d	W_1
LLAMA-3.1-8B	5.71	+0.243	+0.162	0.243	-0.853	-0.580	0.863
QWEN3-14B	5.53	+0.175	+0.118	0.185	-0.935	-0.648	0.945
GEMMA-3-12B	5.61	+0.152	+0.103	0.162	-0.870	-0.596	0.880
GPT-4O-MINI	5.48	+0.098	+0.067	0.108	-0.685	-0.473	0.695
LLAMA-3.3-70B	5.39	+0.082	+0.055	0.092	-0.610	-0.425	0.620
DEEPSEEK-V3	5.42	+0.065	+0.044	0.075	-0.545	-0.388	0.555
GPT-4.1	5.27	+0.020	+0.021	0.060	-0.360	-0.346	0.370

Table 15: Judge-level effects in the broad study (strict, non-CoT; 2,700 nested-test questions per judge). Favorable values average five selected cue families; unfavorable values average all seven. Columns report mean shifts, pooled-SD Cohen's d, and Wasserstein distances; they are not a matched seven-versus-seven polarity comparison. Aggregation: Appendices B.5 and C.2.

Prompt variation. The GPT-4.1 panel in Table 16 averages all seven cue families for each polarity. Unfavorable perturbations produce larger mean shifts (Cohen's $d \in [-0.58, -0.35]$), while favorable aggregates remain near zero ($d \approx 0$). The pattern appears with and without CoT and under both grading rubrics. The strict rubric lowers the baseline mean from $\sim$8.5 to $\sim$5.5; a weak favorable aggregate is therefore also observed away from the upper end of the score scale.

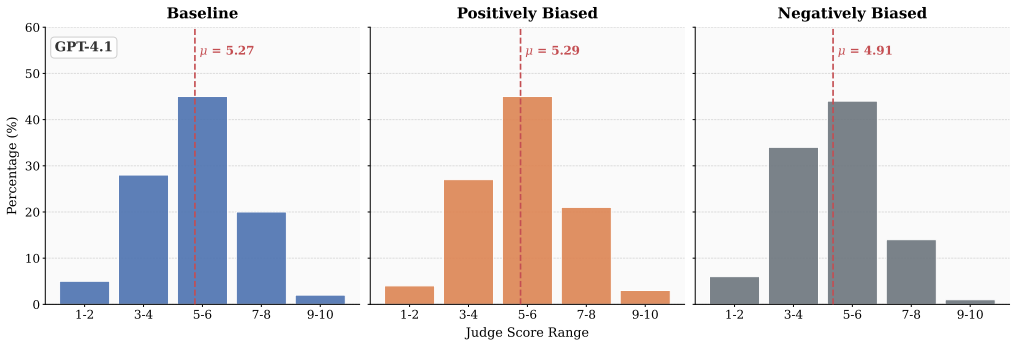
CoT	**Strict**	$\bar{s}_{\text{base}}$	**Favorable (7)**			**Unfavorable (7)**		
			$\Delta\bar{s}$	d	W_1	$\Delta\bar{s}$	d	W_1
✔	✘	8.71	-0.055	-0.043	0.085	-0.855	-0.582	0.855
✘	✘	8.46	-0.010	-0.008	0.080	-0.745	-0.486	0.755
✔	✔	5.72	+0.025	+0.032	0.095	-0.350	-0.382	0.370
✘	✔	5.27	+0.020	+0.021	0.060	-0.360	-0.346	0.370

Table 16: GPT-4.1 responses across the four prompt configurations in the broad study. Strict denotes the grading rubric, not an output-format filter. Both polarities aggregate all seven cue families. Columns report the baseline mean, mean score change, pooled-SD Cohen's d, and Wasserstein distance. Polarity magnitudes are compared descriptively under the statistical procedure in Appendix B.5.

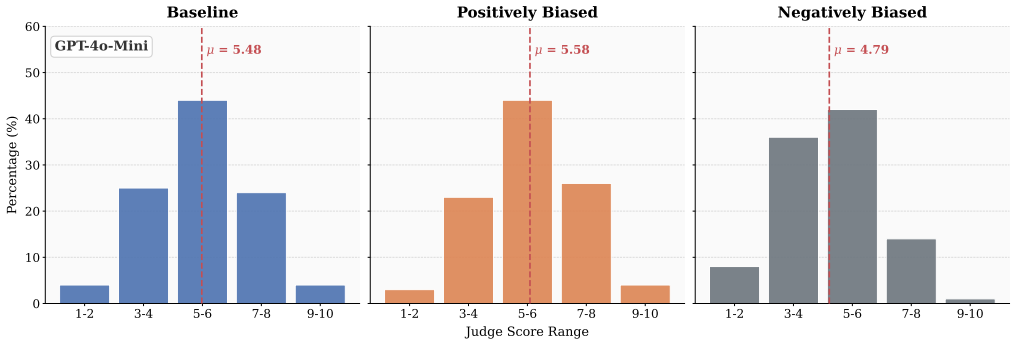
Negative-cue sensitivity appears across judges, while the GPT-4.1 prompt comparisons and full-pool family results show how polarity and domain shape its magnitude. We use asymmetry descriptively for these response patterns under their stated constructions.

Score distributions. Figure 3 shows six judges under the *strict, non-CoT* configuration on the nested test population. Their selected favorable-cue distributions remain close to baseline, while

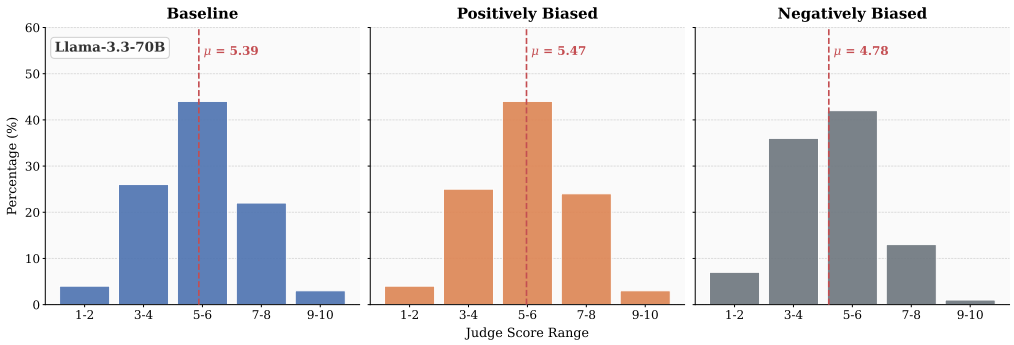
their all-seven-family unfavorable distributions shift leftward. The unfavorable aggregate W_1 ranges from 0.370 on GPT-4.1 to 0.945 on QWEN3-14B (Table 15), reflecting different magnitudes of reassessment across judges.



(a) GPT-4.1

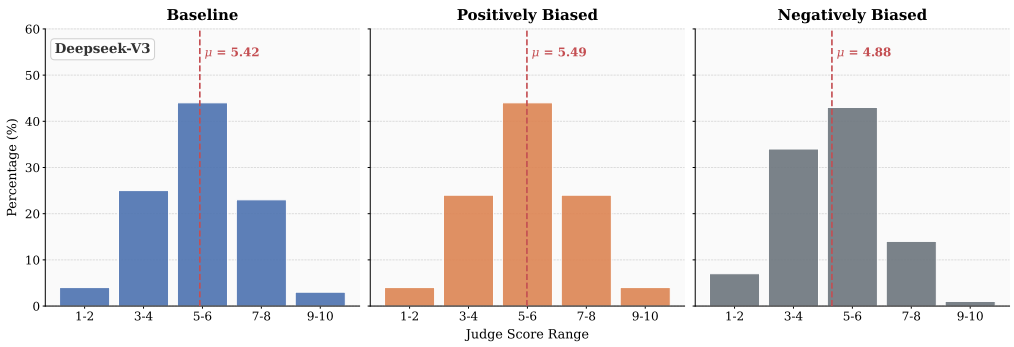


(b) GPT-4O-MINI

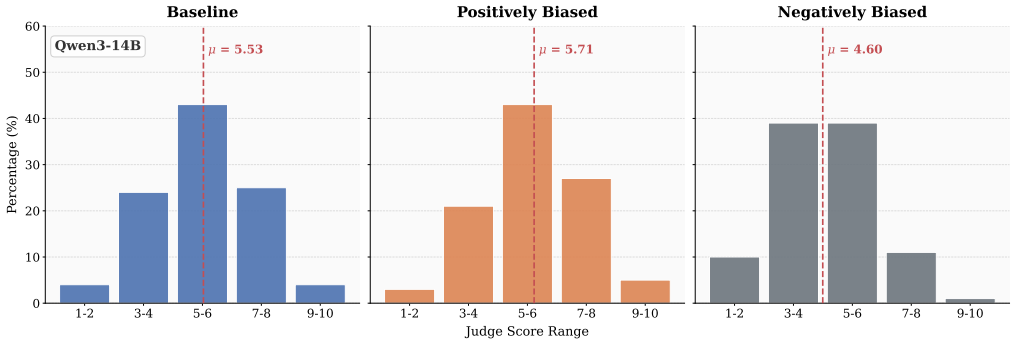


(c) LLAMA-3.3-70B

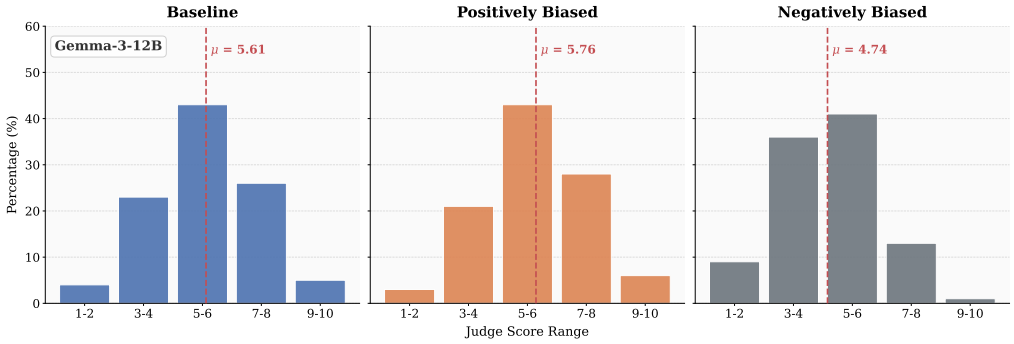
Figure 3: Score distributions for six judges under a strict grading rubric without CoT, using the 2,700-question nested-test population. Favorable curves aggregate the five selected cue families and unfavorable curves all seven, as in Table 15; they are not a family-matched polarity comparison. Unfavorable distributions shift leftward, while selected favorable distributions remain closer to the baseline.



(d) DEEPSEEK-V3



(e) QWEN3-14B



(f) GEMMA-3-12B

Figure 3 (continued).

D CONTROLLED STUDY: DATA AND MEASUREMENT

D.1 QUESTION SAMPLING AND BASELINE ANSWERS

The white-box cohort uses GSM8K (Cobbe et al., 2021), CommonsenseQA (Talmor et al., 2019), and ARC-Challenge (Clark et al., 2018). We sample 400 questions per benchmark with seed 42, using GSM8K's main/test split, CommonsenseQA's validation split, and ARC-Challenge's test split. These source splits identify the published question pools. Our study separately assigns 240 questions per benchmark to training, 80 to development, and 80 to testing before judge inference. All variants of an answer inherit its question's assignment.

Sampling uses the dataset commits in Table 17. Duplicate question stems are removed, and normalized text matching excludes a fixed list of 350 previously used questions. The study's partitions are formed from the remaining pool.

Local PHI-4 (Abdin et al., 2024) generates one candidate answer per question. It receives the question and, for multiple-choice tasks, the listed options, but not the reference answer. Generation is greedy with a 384-token maximum. The prompt requests brief reasoning followed by an explicit final line: `Final answer: <number>` for GSM8K or `Final answer: <LABEL>. <exact option text>` for the two multiple-choice tasks. Parsing requires agreement between the label and option text, allowing only a single terminal sentence period to be added or omitted; the generated response remains unchanged.

Table 18 gives the literal generator instructions. Multiple-choice option text is first normalized by collapsing internal whitespace, then appended to the question under `Options:`, with one `LABEL. text` option per line. "Exact option text" refers to this displayed text. The generator receives one user message through its native chat template.

Well-formed wrong answers remain in the cohort. Outputs that reach the token limit or fail the final-field format are replaced from a deterministically shuffled reserve fixed before generation. A replacement inherits the rejected question's split assignment. This acceptance rule depends on generation completion and final-field format, independently of judge scores; system-level errors do not trigger question replacement.

Reference answers are consulted only after generation to check the final field. A matching numeric answer or matching option establishes final-answer correctness for the content-control eligibility rule; it does not certify every reasoning step. The framing comparison remains within the same stored answer, so both correctly and incorrectly answered questions contribute to that evaluation.

Cohort and control eligibility. Collection accepted 1,200 answers from 1,323 generation attempts. The 123 excluded attempts consist of 116 final-field format failures and seven token-limit terminations. The predetermined reserves filled the affected question slots before any judge scoring. Table 19 reports the accepted population and the final-field reference matches that determine content-control eligibility. All 98 reference-mismatching answers remain in the framing and neutral-metadata evaluations.

D.2 FRAMING, NEUTRAL METADATA, AND CONTENT CONTROLS

All metadata is appended after two newline characters, leaving the stored answer body bit-identical. Bandwagon uses a question-specific percentage $u \in \{70, \ldots, 90\}$ determined by the fixed seed and question identity. The same u is used in both polarities and both wording templates for that question. Refinement changes a claim about review status. The favorable/unfavorable names indicate the intended framing, not an imposed score outcome.

Neutral metadata retains the family's header and bracketed insertion structure but substitutes an archive-code statement. Its digit string is extended deterministically until the full native-chat input has the same token count as the corresponding unfavorable input. Token counts are computed by the judge's tokenizer on the complete prompt, including the question and answer, without running the judge. The matching rule is applied separately to templates A and B. Favorable-input lengths are recorded but are not forced to equal their unfavorable or neutral counterparts.

Both judges use the same question–answer pairs, framing wording, content corruptions, and split assignments. Neutral archive codes are constructed separately with each judge's tokenizer to match

Source	Repository commit
GSM8K	`740312add88f781978c0658806c59bc2815b9866`
CommonsenseQA	`94630fe30dad47192a8546eb75f094926d47e155`
ARC	`210d026faf9955653af8916fad021475a3f00453`

Table 17: Immutable source versions used to build the independent white-box question pool. Published source split names and study split assignments are specified separately in the text.

Common prefix

```
Answer the following question. Explain your reasoning briefly, then give the final
answer.

{question}

{task-specific final instruction}
```

GSM8K final instruction

```
End with exactly one final nonempty line in this format:
Final answer: <number>
Replace <number> with the numerical answer, without units or extra commentary.
```

Multiple-choice final instruction

```
Choose one of the options shown in the question. End with exactly one final
nonempty line in this format:
Final answer: <LABEL>. <exact option text>
Copy the chosen option label and its text exactly; do not add commentary after this
line.
```

Table 18: Literal user-message construction for the local answer generator. No reference answer, correctness flag, judge score, or framing metadata is supplied.

complete-prompt lengths in its native chat format. Thus neutral text can differ between models, while the answer bodies and non-neutral transformations are shared.

Template A is used for direction fitting and development selection, and for a same-wording test on new questions. Template B is reserved for those test questions and changes the metadata wording. Its inputs have no role in fitting the clean-score reference, estimating framing or neutral directions, or selecting intervention strengths.

The content control is generated only when the baseline final-answer field matches the source reference. For GSM8K, the numeric field is replaced by its exact decimal value plus one. For multiple-choice questions, a different option is chosen using the question-specific seed, and both its label and text replace the final field. The original convention concerning a single terminal period is preserved whenever the replacement option permits it. Text outside the parsed final-field span remains byte-identical. This manipulation introduces a known final-answer error, including any resulting inconsistency with the retained reasoning.

Each accepted answer has one clean input and six template-A metadata variants. The 240 test answers additionally have six template-B variants. The 1,102 eligible answers each have one content-altered input, yielding 10,942 inputs per judge. These are repeated evaluations of 1,200 shared questions. Their number is fixed by construction, independently of score validity and response direction.

D.3 Judge Inference and Activation Coordinates

The white-box judges are Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-4B-Instruct-2507 (Qwen Team, 2025). Qwen is pinned to checkpoint commit `cdbee75f17c01a7cc42f958dc650907174af0554`. Each evaluation supplies the question and candidate answer in one user message, applies the checkpoint's native chat template once, and requests an integer score from 1 to 10 without chain-of-thought. The instruction and answer delimiters are

Benchmark	Attempts	Train	Dev	Test	All
GSM8K	406	227/240	77/80	77/80	381/400
CommonsenseQA	502	201/240	73/80	64/80	338/400
ARC-Challenge	415	232/240	78/80	73/80	383/400
Total	1,323	660/720	228/240	214/240	1,102/1,200

Table 19: Actual white-box collection counts. Except for the attempts column, each cell gives final-field reference matches / accepted answers. Reference matches determine eligibility for content corruption; they do not assert that every reasoning step is correct.

Condition	Template A	Template B
Bandwagon+	[Peer review: $u\%$ of reviewers rated this answer as high quality.]	[Reviewer consensus: $u\%$ of reviewers judged the answer to be high quality.]
Bandwagon−	[Peer review: $u\%$ of reviewers rated this answer as low quality.]	[Reviewer consensus: $u\%$ of reviewers judged the answer to be low quality.]
Refinement+	[Review status: This answer has been reviewed.]	[Answer review: The status of this answer is reviewed.]
Refinement−	[Review status: This answer has not been reviewed.]	[Answer review: The status of this answer is not reviewed.]
Neutral Bandwagon	[Peer review: Archive code k.]	[Reviewer consensus: Archive code k.]
Neutral Refinement	[Review status: Archive code k.]	[Answer review: Archive code k.]

Table 20: Literal metadata for the white-box study. The answer body is unchanged. Percentage u is shared within question; archive code k is a deterministic digit string chosen to match its family's unfavorable input in complete judge-prompt token count. Template A appears in all splits; B appears only on test questions.

shown in Table 21. Both models generate greedily over their unrestricted vocabularies, with no forced score tokens, score clipping, or output-dependent retries. The generation budget is eight tokens. Rating extraction and strict adherence to the requested output format are distinct outcomes, defined in Appendix D.4.

We observe raw decoder-block outputs at zero-based indices 15, 23, and 31 in Llama, and 17, 26, and 35 in Qwen. These are the half-depth, three-quarter-depth, and final blocks of the respective 32- and 36-block models. The intervention blocks are fixed at 23 and 26. Final-block observations precede each model's final RMS normalization, so projections and additions consistently use raw block coordinates. A hook locates the last unmasked input position in each padded sequence. It records that state during prefill and, when enabled, adds the intervention there once. Later decoding steps are not modified.

For Llama, the first generated-token logits for the ten bare integer tokens supply the targets for its logit-augmented score reference. These logits are measured separately from the generated rating. Qwen represents score 10 with two tokens, so the ten-label first-token regression is confined to Llama; the common reference uses scalar ratings and the high–low contrast in both judges. State availability and textual score validity are assessed separately: a successfully computed finite state can precede a completion that fails the score-format rule.

Inference runs on one NVIDIA RTX PRO 6000 Blackwell GPU with 96 GB memory, using PyTorch 2.11, CUDA 13.0, and Transformers 5.13. The judge uses FP32 with TF32 disabled; the answer generator uses BF16. Scoring and interventions share a fixed input order, 128-record shard boundaries, and batch size 16. Inputs exceeding 2,048 tokens are flagged without truncation. Numerical checks before acquisition compare generated outputs, parsing states, and raw activations across mixed-length batches; zero-strength interventions reproduce no-intervention inference on the same inputs.

White-box judge: strict, non-CoT user message

```
You are a **very strict** judge. Given a question and a candidate answer, rate the
quality of the candidate answer from 1 to 10 — **only truly exceptional** answers
deserve a score of 7 or higher.

[The start of the question]
{question}
[The end of the question]

[The start of the candidate answer]
{answer}
[The end of the candidate answer]

Output Format:
Output only one single integer — the score. Do not output anything else. This
number is the final score, and it must be the only thing in your response.
```

Table 21: User-message content for the white-box judge. Asterisks are literal prompt characters; placeholders are replaced by the question and candidate answer, including any appended metadata. The native chat template adds the model's conversation tokens.

D.4 Rating Extraction and Output-Format Compliance

The primary endpoint is the initially emitted rating. An extractable rating consists of an ASCII integer from 1 to 10 on the first physical line, with optional ASCII spaces or tabs and an observed line-break or end-of-sequence boundary. Leading empty lines, numeric prefixes followed by other text on that line, and out-of-range values fail this rule. A bare numeral at the token limit without a completed line is also invalid. Runtime failures and nonfinite states remain invalid.

A completed integer line can be followed by additional text, including a suffix that reaches the eight-token limit. This endpoint measures the initial rating, not the eventual conclusion of a longer completion. The separate strict-format endpoint requires end-of-sequence termination and a complete whitespace-stripped response consisting of exactly one in-range integer. Additional explanation, a decimal, or token-limit termination fails that endpoint. We report extraction, strict compliance, suffix occurrence, and token-limit termination separately; the 93% development constraint concerns rating extraction.

For Llama, all 10,942 unsteered inputs yield an extractable first-line rating. Of these, 7,038 satisfy the strict format; the other 3,904 complete a rating line and then continue to the token limit. Among its 1,200 clean evaluations, strict compliance is 494/720 on training questions, 164/240 on development questions, and 162/240 on test questions. These format counts are distinct from the primary score-fitting and direction-discovery counts. The common-reference results give both models' policy-specific coverage alongside their score responses.

Scoring policy and study timing. For Llama, first-line scoring was adopted after inspecting unsteered outputs from all three splits, before steering search or test interventions. Every score-dependent regression, direction, and matching set was refitted using the first-line labels, with the same question splits, raw states, prompts, model, and inference configuration. Development outcomes then selected interventions before test steering. The common-reference Llama comparison and complete response-curve follow-up were specified after the logit-augmented Llama results and are retrospective analyses of that cohort. In contrast, the Qwen replication fixed the scoring policy, common reference, and intervention protocol before acquiring Qwen outcomes. Strict-format sensitivity uses the primary directions, strengths, inputs, and completions in both models. Its scoring policy determines validity under each analysis's stated population and loss rules (Appendix E.5).

E CONTROLLED STUDY: ESTIMATION AND EVALUATION

E.1 CLEAN-SCORE REGRESSION AND PROJECTION

Each observed block has its own clean-score reference, fitted from unframed training evaluations with extractable primary ratings. Unless explicitly labeled strict-format, score validity in this appendix refers to that first-line endpoint. The study assigns 720 questions to training and 240 to development, with usable score counts recorded for each judge. Let $H \in \mathbb{R}^{n\times d}$ contain the valid training states. We subtract the training mean μ and divide every coordinate by the same scalar τ, the root mean square of the centered entries of H. This scalar scaling preserves the Euclidean geometry used for projection. Targets are centered by their training means. Ridge regression solves

$$B_\lambda = \arg\min_B \|ZB - Y\|_F^2 + \lambda\|B\|_F^2, \qquad Z = (H - \mathbf{1}\mu^\top)/\tau. \tag{5}$$

The common reference uses the scalar rating as target Y. Regularization is selected by development mean squared error over the fixed sequence $\{0.001, 0.01, 0.1, 1, 10, 100\}$. An exact tie selects the first value. Predictions add back the training target mean. Coefficients are divided by τ before constructing the reference in raw activation coordinates.

The high–low reference uses the global 25th and 75th percentiles of valid clean training scores, including observations at the respective cutoffs. Within each benchmark, the larger eligible arm is subsampled to the size of the smaller arm using the fixed seed. The direction is the pooled high-score mean minus the pooled low-score mean. The arm identities are the same at all observed blocks. A benchmark with no support in one arm contributes neither arm; equal global cutoffs make this reference unavailable. Counts, cutoffs, and selected score ranges accompany the fitted reference.

Let w_l denote the scalar coefficient in raw coordinates and m_l the high–low difference. We normalize the nonzero vectors, then use their singular-value decomposition to obtain Q_l. Singular values above $\max(10^{-10}\sigma_1, 10^{-12})$ define its rank. The common reference consequently has rank at most two in each model. Test score MAE, Spearman correlation, and R^2 assess prediction by the fitted scalar regression. Its orientation and the high–low support are reported separately because a predictive coefficient and a between-group contrast need not have the same direction.

Logit-augmented Llama reference. A complementary Llama analysis adds a multi-output ridge regression to centered integer-token logits. Within each input, the mean of the ten logits is subtracted, so the target describes score-token contrasts rather than a common additive offset. Regularization uses the same development grid, selected independently of the scalar regression. After fitting, the mean coefficient row is removed to eliminate numerical common-logit components. The resulting raw-coordinate coefficient matrix W_l has row-space rank at most nine. The expanded reference is $\mathcal{U}_l^{\text{logit}} = \text{span}\{w_l, m_l, \text{rows}(W_l)\}$ and has rank at most eleven. This analysis is confined to Llama, whose ten integer ratings have bare single-token representations (Appendix D.3).

Both Llama references use the same scalar coefficient and high–low contrast. Changing the reference changes the projections and component directions while leaving the unprojected framing and content contrasts fixed. Each unit component receives development selection under its own reference. A residual is orthogonal to the declared fitted span; neither span exhausts all scoring information. The timing of the two-model comparison and scoring-policy specification is given in Appendix D.4.

E.2 DISPLACEMENT GEOMETRY AND RESPONSE GROUPS

The effective group consists of valid baseline–framing score pairs satisfying $\epsilon_c \Delta s_i^c > 0$. Unchanged-score pairs satisfy $\Delta s_i^c = 0$, and opposite-response pairs satisfy $\epsilon_c \Delta s_i^c < 0$. Invalid pairs are counted separately. The primary mean direction uses effective training pairs with activations at both endpoints. The group definition uses the score-response sign, with no filter on displacement magnitude or cosine. A condition with fewer than 30 effective training pairs is labeled descriptive.

For a matrix D whose rows are paired displacements, let $D_0 = D - \mathbf{1}\bar{\Delta}h^\top$ be its centered version. The score-reference energy fractions are

$$E_{\text{total}} = \frac{\|DQ_l\|_F^2}{\|D\|_F^2}, \qquad E_{\text{centered}} = \frac{\|D_0 Q_l\|_F^2}{\|D_0\|_F^2}. \tag{6}$$

Residual fractions are their complements. The total-energy measure includes the mean displacement, whereas the centered measure concerns variation around that mean. For the unit mean direction itself, $\|Q_l^\top v_l^c\|_2^2$ is its score-parallel energy fraction. The latter describes the mean orientation, while the two matrix measures describe the population of displacements.

Covariance spectra are computed for D, DQ_l, and $D - DQ_lQ_l^\top$, centering each matrix before eigendecomposition. The cumulative shares for 1, 3, 5, and 10 principal components divide by that matrix's own total centered energy. A residual spectrum describes the concentration of retained variation; its energy fraction separately measures how much variation survives projection. Spectra are undefined for zero-energy matrices or fewer than two pairs.

Counts and geometry are evaluated separately for effective, unchanged-score, and opposite-response groups in each condition and split. Held-out group projections use the unit mean direction fitted on effective training pairs. We summarize the projected displacements by their mean, standard deviation, quantiles, and mean cosine with this fixed orientation. Test covariance spectra describe variation among test pairs; intervention directions are fitted on training pairs. The training-set first principal component, sign-aligned to the training mean, supplies a supplementary geometric orientation, while the primary interventions use the mean direction.

The parallel and residual components sum to the full unit mean direction. Raw components retain their projected magnitudes and use the full direction's selected strength. A component whose norm exceeds 0.001 is also normalized to unit length and receives independent development selection over the common grid. The raw comparison measures effects at the sizes present in the full vector; the unit comparison measures control by each orientation under its own selected strength.

E.3 Exact Score-Change Matching

Matching is performed separately for Bandwagon− and Refinement−, using template A, and separately within each study split. Both arms draw from the same parent population: baseline answers whose final field matches the benchmark reference. A framing or content pair is eligible when both scores are valid, both activation records exist, and the altered score is lower than its baseline. Its stratum is the exact tuple $(\text{benchmark}, s_{\text{base}}, s_{\text{base}} - s_{\text{altered}})$.

Within a stratum, questions present in both arms are paired first. The remaining members of each arm are shuffled using the seed and the split/condition identity, then matched without replacement until the smaller arm is exhausted. A match can therefore link different parent questions, while every displacement remains a within-answer difference. The two selected arms have identical empirical benchmark, baseline-score, and score-drop distributions. Their mean displacements, covariance spectra, and mean-direction cosine are computed in the same block coordinates.

Training matches define separate matched-framing and matched-content mean directions; their intervention eligibility follows the support rule below. Eligible and selected counts, benchmark composition, same-question counts, and each arm's coverage characterize the matched population. A primary matched comparison requires at least 80 training matches and 50 test matches, each spanning at least two benchmarks. Comparisons below this support requirement are descriptive and use the same exact matching rule. Outcome-group prevalence in the complete framing population is reported separately from matched-set coverage.

E.4 Intervention Grid and Development Selection

The intervention block l_* is fixed at raw decoder-block index 23 for Llama and 26 for Qwen. For a unit direction v, the intervention strengths are

$$\alpha = r \ \text{median}_{i \in \mathcal{D}_{\text{train}}^{\text{clean}}} \|h_{l_*}(x_i)\|_2, \qquad r \in \{0, \pm 0.005, \pm 0.01, \pm 0.02, \pm 0.04, \pm 0.08, \pm 0.16\}. \tag{7}$$

The median uses valid clean training states for the evaluated judge. Both signs are searched for every direction. The common-reference forward study covers Bandwagon− and Refinement−, comparing the full mean, its unit parallel and residual components, the clean high–low and scalar-ridge directions, its family's neutral mean, and five random unit directions. Matched-framing and matched-content means are included when the matching-support rule is satisfied. Neutral means use all training neutral displacements with both activation endpoints, without an outcome filter. Random vectors

are independent isotropic Gaussian draws normalized to unit length, with seeds 101–105 combined with the block index. Each draw receives the entire grid and the same selection rule, and all five are reported.

Selection objectives. Forward selection uses all 240 clean development answers. For the two unfavorable cues its objective is the mean decrease over valid baseline/intervention pairs. The logit-augmented Llama study additionally evaluates favorable conditions and performs reverse selection. Reverse search uses all 240 template-A development inputs of the target unfavorable condition, including unchanged and opposite responses. It minimizes the mean absolute difference from each answer's clean score, assigning loss nine whenever either required score is invalid. Both objectives evaluate the complete development population rather than its effective subset.

A candidate is feasible only if at least 93% of its outputs have extractable first-line ratings and Spearman correlation with the clean baseline is at least 0.85. Strict integer-only compliance is reported separately and does not select the configuration. Correlation uses complete baseline/intervention pairs and average ranks for ties; fewer than two pairs or a constant rank vector gives an undefined correlation and fails the constraint. Among feasible candidates, the best objective wins, followed by smaller $|\alpha|$ and then smaller signed α. Zero is included in the grid and can be selected. If no candidate is feasible, we report zero as an infeasible reference.

Held-out evaluation. Development selection fixes the direction and coefficient before test steering. Each selected common-reference configuration is evaluated on clean test answers. Raw parallel and residual components inherit the full mean's selected α; its development metrics describe the full direction, not component-specific feasibility. The logit-augmented Llama reverse configurations are evaluated on four populations using the same direction and strength: framed template A, framed template B, clean answers, and eligible content-altered answers. This paired application relates recovery of framed ratings to changes on inputs without that framing.

E.5 Score Effects, Paired Recovery, and Uncertainty

Common-reference forward endpoint. For scoring policy p, the reference population R_p comprises clean test questions with an originally valid rating under that policy. This set is fixed before intervention. Each member contributes the downward change $s_{\mathrm{base},p} - s_{\mathrm{steer},p}$ when the steered rating is valid, and -9 otherwise. We report the mean of these contributions alongside the valid-pair mean, total question count, fixed-reference count, and complete-pair count. Initially invalid references reduce coverage and are excluded from R_p. Primary and strict-format results use their own R_p with the same primary-selected direction and strength. At zero strength, the change is zero under either policy.

Paired recovery on fixed references. The response-curve population additionally requires a valid original score for the evaluated input role: $R_p = \{i : s_{\mathrm{base},p}, s_{\mathrm{original},p} \text{ valid}\}$. On that fixed set, recovery is the reduction in absolute error relative to the question's clean score. A later invalid steered score receives absolute-error loss nine. Thus the mean gain is the mean of $|s_{\mathrm{original}} - s_{\mathrm{base}}| - L_{\mathrm{steer}}$, where $L_{\mathrm{steer}} = |s_{\mathrm{steer}} - s_{\mathrm{base}}|$ for a valid score and nine otherwise. Complete-valid-triple gains and their coverage are also reported. This fixed-reference definition gives zero gain at zero strength; missing original references reduce its population size.

Movement toward baseline, worsening, and unchanged error compare the new penalized loss with the original absolute error on the fixed reference set. Exact recovery uses its originally changed subset and counts subsequent invalid outputs as not recovered. Overshoot uses valid triples in that originally changed subset and requires strictly opposite signed errors; landing exactly on baseline is recovery, not overshoot. These movement fractions supplement the MAE endpoint and retain their separate denominators.

Collateral change and content gaps. For clean and content-altered inputs, collateral change is $|s_{\mathrm{steer}} - s_{\mathrm{original}}|$ on the fixed original-reference support, with loss nine for a later invalid score. A valid-pair mean accompanies it. The fixed-clean gap is $s_{\mathrm{base}} - s_{\mathrm{content,steer}}$, while the jointly steered gap is $s_{\mathrm{clean,steer}} - s_{\mathrm{content,steer}}$. Both are compared with the original gap $s_{\mathrm{base}} - s_{\mathrm{content}}$. The joint view measures whether the same intervention preserves the judge's clean/content separation. Penalized absolute gap changes assign nine if a required steered endpoint is invalid; for the joint gap

this is a specified cost, not a worst-case bound. Valid-endpoint summaries and their counts remain separate.

Framed results are accompanied by effective, unchanged-score, and opposite-response subgroups defined from unsteered scores. A framed template uses its own response for subgroup membership. Clean/content collateral subgroups use the target condition's template-A response on that question. Membership is fixed before intervention and is not reassigned according to recovery. Spearman and marginal Wasserstein-1 distance use complete score pairs and report their coverage; Wasserstein distance remains a secondary distributional summary.

Logit-augmented selected-configuration endpoint. The Llama selected-configuration tables in Appendix J.1 use an all-input population. An invalid original baseline or framed reference contributes -9 to conservative gain, and either invalid endpoint contributes nine to collateral cost. Strict-format sensitivity can therefore have nonzero penalties even at zero strength. These all-input estimates and their valid-only counterparts answer a different population question from the fixed-reference endpoints above; their labels identify the applicable convention.

Uncertainty. We use 2,000 question-cluster bootstrap resamples with seed 42, stratified by benchmark. Within each reported population or contrast, each resampled question carries all of its associated records, with multiplicity, and method contrasts use common question identities. In the common-reference forward comparison, draws resample the query population and then apply each scoring policy's fixed original-validity mask. The 95% percentile intervals condition on fitted directions, matched sets, and development-selected configurations; fitting, matching, and strength selection are not repeated within a bootstrap draw. Matched-arm intervals retain the question clusters in both arms, including for cross-question matches. Thus the two IDs in a match remain distinct observational units.

The logit-augmented selected recovery analysis classifies collateral effects as small only when the upper confidence bound on conservative collateral change is below 0.25 score points. It also requires a complete panel, valid endpoints, at least two questions, all requested bootstrap draws, and a feasible development selection. A nonsignificant difference from zero alone does not satisfy this criterion. The complete response curves show all sampled strengths with pointwise intervals, independently of that selection outcome. They characterize a finite grid rather than a test-selected operating point or a global control frontier.

F Controlled Responses and Coverage

Llama-3.1-8B-Instruct and Qwen3-4B-Instruct-2507 evaluate the same question–answer pairs and study partitions. Their paired ratings determine the response populations used in the representation and intervention comparisons. The Llama populations also support the logit-augmented analysis in Appendix I; changing the clean-score reference changes the decomposition, not those populations.

F.1 Responses, Templates, and Output Coverage

Framing, neutral metadata, and final-answer errors produce different response frequencies on the shared answers. We report training and development responses together, followed by the held-out comparison. Each model's mean rating is accompanied by its paired score change and numbers of decreases, unchanged ratings, and increases. Template A is evaluated in every split; template B uses the same test questions with alternative wording.

		Llama: first-line			Llama: strict			Qwen: both policies		
Split/condition/T	N	n	**Mean** Δs	$\downarrow / = / \uparrow$	n	**Mean** Δs	$\downarrow / = / \uparrow$	n	**Mean** Δs	$\downarrow / = / \uparrow$
Dev/BW −/A	240	240	3.921 -3.725	240/0/0	120	4.122 -3.950	120/0/0	240	4.558 -5.000	236/4/0
Dev/BW +/A	240	240	7.283 -0.362	53/186/1	157	7.925 -0.134	20/136/1	240	9.450 -0.108	56/166/18
Dev/Clean	240	240	7.646 0.000	0/240/0	164	8.043 0.000	0/164/0	240	9.558 0.000	0/240/0
Dev/Final-answer error	228	228	4.184 -3.461	220/8/0	104	4.137 -3.990	102/2/0	224	2.576 -7.036	216/8/0
Dev/Neutral BW/A	240	240	7.004 -0.642	84/156/0	148	7.920 -0.169	22/126/0	240	9.429 -0.129	23/212/5
Dev/Neutral RF/A	240	240	7.296 -0.350	50/190/0	160	7.896 -0.138	19/141/0	240	9.454 -0.104	16/221/3
Dev/RF −/A	240	240	7.554 -0.092	23/211/6	161	7.958 -0.087	14/146/1	240	9.508 -0.050	10/226/4
Dev/RF +/A	240	240	7.412 -0.233	35/205/0	161	7.945 -0.124	17/144/0	240	9.558 0.000	3/232/5
Train/BW −/A	720	720	3.975 -3.606	715/5/0	355	4.096 -3.961	355/0/0	718	4.331 -5.039	700/18/0
Train/BW +/A	720	720	7.244 -0.336	157/555/8	458	7.913 -0.177	73/381/4	718	9.326 -0.035	156/501/61
Train/Clean	720	720	7.581 0.000	0/720/0	494	8.049 0.000	0/494/0	718	9.376 0.000	0/718/0
Train/Final-answer error	660	660	4.056 -3.524	627/33/0	279	4.010 -4.151	273/6/0	644	2.512 -6.997	619/23/2
Train/Neutral BW/A	720	720	6.926 -0.654	264/456/0	446	7.856 -0.253	90/356/0	718	9.206 -0.160	80/625/13
Train/Neutral RF/A	720	720	7.218 -0.362	161/558/1	476	7.843 -0.164	70/405/1	718	9.235 -0.131	59/647/12
Train/RF −/A	720	720	7.472 -0.108	71/634/15	489	7.925 -0.094	44/441/4	718	9.341 -0.025	36/660/22
Train/RF +/A	720	720	7.382 -0.199	99/615/6	486	7.895 -0.126	54/430/2	718	9.382 0.017	22/671/25

Table 22: Training and development responses of the two judges. P/S denote first-line and strict whole-response scoring; both policies coincide for the displayed Qwen conditions. N counts queried inputs and n valid clean–condition pairs. Mean rating uses condition-valid outputs; Δs is the paired condition-minus-clean change. The two-line mean/Δs cells preserve these distinct summaries. Counts $\downarrow / = / \uparrow$ are paired decreases, unchanged scores, and increases. T denotes wording template; BW/RF denote Bandwagon/Refinement. Entries are descriptive point estimates.

The held-out responses preserve the contrast between the two unfavorable cues. Bandwagon decreases 239 of 240 Llama ratings and 234 of 240 Qwen ratings on template A, whereas Refinement leaves 208 and 222 ratings unchanged. Alternative wording changes the size of the Bandwagon response differently across judges: its mean score change moves from -3.592 to -2.996 in Llama and from -5.167 to -6.900 in Qwen. The shared response pattern therefore coexists with model-dependent sensitivity to wording.

		Llama: first-line			**Llama: strict**			**Qwen: both policies**		
Split/condition/T	N	n	**Mean** Δs	$\downarrow / = / \uparrow$	n	**Mean** Δs	$\downarrow / = / \uparrow$	n	**Mean** Δs	$\downarrow / = / \uparrow$
			4.013			4.115			4.446	
Test/BW $-$/A	240	240	-3.592	239/1/0	114	-3.842	113/1/0	240	-5.167	234/6/0
			4.608			4.720			2.712	
Test/BW $-$/B	240	240	-2.996	239/1/0	153	-3.176	152/1/0	240	-6.900	238/2/0
			7.279			7.930			9.429	
Test/BW $+$/A	240	240	-0.325	52/185/3	155	-0.168	25/128/2	240	-0.183	58/172/10
			7.463			7.971			9.208	
Test/BW $+$/B	240	240	-0.142	37/184/19	160	0.013	16/128/16	240	-0.404	119/111/10
			7.604			8.056			9.613	
Test/Clean	240	240	0.000	0/240/0	162	0.000	0/162/0	240	0.000	0/240/0
			4.023			3.972			2.591	
Test/Final-answer error	214	214	-3.565	204/9/1	91	-4.242	90/1/0	208	-7.120	200/8/0
			7.083			7.967			9.392	
Test/Neutral BW/A	240	240	-0.521	72/168/0	151	-0.159	23/128/0	240	-0.221	23/216/1
			7.058			7.936			9.442	
Test/Neutral BW/B	240	240	-0.546	76/164/0	153	-0.150	22/131/0	240	-0.171	19/219/2
			7.267			7.869			9.392	
Test/Neutral RF/A	240	240	-0.338	50/190/0	161	-0.118	19/142/0	240	-0.221	24/214/2
			6.996			7.929			9.483	
Test/Neutral RF/B	240	240	-0.608	81/159/0	153	-0.157	24/129/0	240	-0.129	12/227/1
			7.508			7.976			9.537	
Test/RF $-$/A	240	240	-0.096	24/208/8	160	-0.044	12/144/4	240	-0.075	14/222/4
			7.554			8.018			9.521	
Test/RF $-$/B	240	240	-0.050	20/208/12	160	-0.013	9/145/6	240	-0.092	15/222/3
			7.429			7.946			9.546	
Test/RF $+$/A	240	240	-0.175	31/205/4	161	-0.062	15/142/4	240	-0.067	14/221/5
			7.404			7.969			9.488	
Test/RF $+$/B	240	240	-0.200	35/202/3	156	-0.103	17/138/1	240	-0.125	14/224/2

Table 23: Held-out responses of the two judges. P/S denote first-line and strict whole-response scoring; both policies coincide for the displayed Qwen conditions. N counts queried inputs and n valid clean–condition pairs. Mean rating uses condition-valid outputs; Δs is the paired condition-minus-clean change. The two-line mean/Δs cells preserve these distinct summaries. Counts $\downarrow / = / \uparrow$ are paired decreases, unchanged scores, and increases. T denotes wording template; BW/RF denote Bandwagon/Refinement. Entries are descriptive point estimates.

Rating extraction and whole-response format compliance describe different properties of these outputs. Llama often produces an extractable first-line score followed by additional text; Qwen's extractable scores also satisfy the whole-response rule in these conditions. The coverage comparison keeps these format categories separate from valid clean–condition pairs, which require both ratings. Decoder-limit termination is a further output property and need not coincide with an invalid rating. Appendix D.4 defines the two scoring policies.

		Llama				Qwen			
Split/condition/T	N	**First**	**Whole**	**Suffix**	**Cap**	**First**	**Whole**	**Suffix**	**Cap**
Dev/BW −/A	240	240	131	109	109	240	240	0	0
Dev/BW +/A	240	240	159	81	81	240	240	0	0
Dev/Clean	240	240	164	76	76	240	240	0	0
Dev/Final-answer error	228	228	117	111	111	224	224	0	4
Dev/Neutral BW/A	240	240	150	90	90	240	240	0	0
Dev/Neutral RF/A	240	240	164	76	76	240	240	0	0
Dev/RF −/A	240	240	165	75	75	240	240	0	0
Dev/RF +/A	240	240	163	77	77	240	240	0	0
Test/BW −/A	240	240	131	109	109	240	240	0	0
Test/BW −/B	240	240	189	51	51	240	240	0	0
Test/BW +/A	240	240	157	83	83	240	240	0	0
Test/BW +/B	240	240	173	67	67	240	240	0	0
Test/Clean	240	240	162	78	78	240	240	0	0
Test/Final-answer error	214	214	107	107	107	208	208	0	3
Test/Neutral BW/A	240	240	153	87	87	240	240	0	0
Test/Neutral BW/B	240	240	157	83	83	240	240	0	0
Test/Neutral RF/A	240	240	168	72	72	240	240	0	0
Test/Neutral RF/B	240	240	155	85	85	240	240	0	0
Test/RF −/A	240	240	166	74	74	240	240	0	0
Test/RF −/B	240	240	166	74	74	240	240	0	0
Test/RF +/A	240	240	167	73	73	240	240	0	0
Test/RF +/B	240	240	159	81	81	240	240	0	0
Train/BW −/A	720	720	385	335	335	720	720	0	0
Train/BW +/A	720	720	469	251	251	720	720	0	0
Train/Clean	720	720	494	226	226	718	718	0	0
Train/Final-answer error	660	660	308	352	352	646	646	0	2
Train/Neutral BW/A	720	720	451	269	269	719	719	0	0
Train/Neutral RF/A	720	720	496	224	224	719	719	0	0
Train/RF −/A	720	720	508	212	212	719	719	0	0
Train/RF +/A	720	720	504	216	216	719	719	0	0

Table 24: Unsteered output coverage, with Llama columns followed by Qwen columns. First counts extractable first-line integer ratings; Whole counts whole-response integer-only compliance. Suffix counts nonempty text after the first physical line and Cap counts decoder-limit termination. These categories are separate and need not be mutually exclusive. N is the queried population; the decoder limit remains eight tokens. BW/RF denote Bandwagon/Refinement; T denotes template.

Effective responses are valid nonzero changes in the cue's intended direction. Their domain-specific counts connect the behavioral comparison to the populations used for direction estimation. The abundant Bandwagon-negative responses support a broad matched-source comparison, while Refinement-negative supplies both effective and unchanged responses within the same cue. Favorable-cue support varies substantially across judges and domains, so the table retains each condition's counts rather than pooling polarities.

Split/cue/T	ARC L	ARC Q	CSQA L	CSQA Q	GSM8K L	GSM8K Q
Train/BW +/A	240/3	240/13	240/2	240/34	240/3	240/14
Dev/BW +/A	80/1	80/4	80/0	80/10	80/0	80/4
Test/BW +/A	80/0	80/2	80/2	80/6	80/1	80/2
Test/BW +/B	80/8	80/2	80/8	80/6	80/3	80/2
Train/BW −/A	240/240	240/236	240/240	240/233	240/235	240/231
Dev/BW −/A	80/80	80/78	80/80	80/79	80/80	80/79
Test/BW −/A	80/80	80/80	80/80	80/78	80/79	80/76
Test/BW −/B	80/80	80/80	80/80	80/78	80/79	80/80
Train/RF +/A	240/1	240/8	240/1	240/13	240/4	240/4
Dev/RF +/A	80/0	80/2	80/0	80/3	80/0	80/0
Test/RF +/A	80/1	80/0	80/2	80/3	80/1	80/2
Test/RF +/B	80/0	80/0	80/1	80/1	80/2	80/1
Train/RF −/A	240/25	240/10	240/19	240/21	240/27	240/5
Dev/RF −/A	80/6	80/4	80/7	80/6	80/10	80/0
Test/RF −/A	80/10	80/7	80/2	80/6	80/12	80/1
Test/RF −/B	80/7	80/8	80/2	80/6	80/11	80/1

Table 25: Effective-response support by question domain and judge. Each cell gives all activation-paired inputs/effective responses. Effective means a valid nonzero score change in the cue's intended direction. L/Q denote Llama/Qwen; ARC is ARC-Challenge and CSQA is CommonsenseQA. Train/dev use template A; held-out templates A and B share questions. Empty response groups remain zero.

F.2 Neutral-Metadata Contrasts

The Llama framing-minus-neutral contrast measures the incremental rating response to an evaluative statement relative to neutral metadata on the same answer. Subtracting the two ratings cancels their common clean reference. The neutral variants are length-matched to unfavorable framing; favorable contrasts reuse their corresponding neutral variants without length matching. The paired-policy table places each contrast beside the two component changes and its valid clean/framed/neutral intersection.

Bandwagon-negative produces a substantially larger score decrease than neutral metadata under both policies. On test template A, the framing-minus-neutral estimate is -3.071 under first-line scoring and -3.787 under strict scoring. Refinement-negative instead lowers scores less than its neutral counterpart, giving positive contrasts of 0.242 and 0.075. Thus, the neutral comparison distinguishes a strong evaluative-metadata response from a small overall framing response. Each policy's estimate refers to its own displayed intersection rather than a common population across policies.

Cue	Split	T	Rule	n	Δ_f	Δ_n	$\Delta_f - \Delta_n$
BW−	dev	A	P	240	-3.725 [-3.837, -3.625]	-0.642 [-0.754, -0.550]	-3.083 [-3.196, -2.975]
BW−	dev	A	S	114	-3.965 [-4.061, -3.851]	-0.149 [-0.219, -0.079]	-3.816 [-3.921, -3.711]
BW+	dev	A	P	240	-0.362 [-0.458, -0.283]	-0.642 [-0.754, -0.550]	0.279 [0.196, 0.367]
BW+	dev	A	S	147	-0.129 [-0.190, -0.075]	-0.170 [-0.238, -0.109]	0.041 [0.000, 0.089]
RF−	dev	A	P	240	-0.092 [-0.167, -0.025]	-0.350 [-0.438, -0.271]	0.258 [0.183, 0.338]
RF−	dev	A	S	160	-0.087 [-0.150, -0.031]	-0.138 [-0.200, -0.081]	0.050 [0.013, 0.094]
RF+	dev	A	P	240	-0.233 [-0.312, -0.167]	-0.350 [-0.438, -0.271]	0.117 [0.062, 0.179]
RF+	dev	A	S	160	-0.125 [-0.188, -0.069]	-0.138 [-0.200, -0.081]	0.013 [-0.006, 0.037]
BW−	test	A	P	240	-3.592 [-3.692, -3.496]	-0.521 [-0.613, -0.433]	-3.071 [-3.171, -2.967]
BW−	test	A	S	108	-3.907 [-4.019, -3.787]	-0.120 [-0.194, -0.065]	-3.787 [-3.907, -3.648]
BW−	test	B	P	240	-2.996 [-3.108, -2.879]	-0.546 [-0.633, -0.458]	-2.450 [-2.567, -2.333]
BW−	test	B	S	146	-3.192 [-3.342, -3.027]	-0.151 [-0.212, -0.096]	-3.041 [-3.192, -2.877]
BW+	test	A	P	240	-0.325 [-0.408, -0.237]	-0.521 [-0.613, -0.433]	0.196 [0.121, 0.271]
BW+	test	A	S	151	-0.172 [-0.238, -0.106]	-0.159 [-0.219, -0.099]	-0.013 [-0.079, 0.053]
BW+	test	B	P	240	-0.142 [-0.233, -0.054]	-0.546 [-0.633, -0.458]	0.404 [0.312, 0.496]
BW+	test	B	S	153	0.013 [-0.065, 0.092]	-0.150 [-0.216, -0.092]	0.163 [0.092, 0.235]
RF−	test	A	P	240	-0.096 [-0.171, -0.025]	-0.338 [-0.421, -0.262]	0.242 [0.171, 0.321]
RF−	test	A	S	160	-0.044 [-0.094, 0.006]	-0.119 [-0.169, -0.069]	0.075 [0.037, 0.119]
RF−	test	B	P	240	-0.050 [-0.129, 0.025]	-0.608 [-0.704, -0.517]	0.558 [0.471, 0.654]
RF−	test	B	S	153	-0.026 [-0.078, 0.026]	-0.157 [-0.216, -0.105]	0.131 [0.085, 0.190]
RF+	test	A	P	240	-0.175 [-0.246, -0.108]	-0.338 [-0.421, -0.262]	0.163 [0.104, 0.229]
RF+	test	A	S	161	-0.062 [-0.118, -0.006]	-0.118 [-0.168, -0.075]	0.056 [0.025, 0.093]
RF+	test	B	P	240	-0.200 [-0.275, -0.125]	-0.608 [-0.704, -0.517]	0.408 [0.329, 0.487]
RF+	test	B	S	151	-0.106 [-0.152, -0.059]	-0.159 [-0.212, -0.106]	0.053 [0.020, 0.093]
BW−	train	A	P	720	-3.606 [-3.668, -3.543]	-0.654 [-0.713, -0.592]	-2.951 [-3.015, -2.888]
BW−	train	A	S	330	-3.997 [-4.067, -3.924]	-0.267 [-0.330, -0.203]	-3.730 [-3.806, -3.652]
BW+	train	A	P	720	-0.336 [-0.388, -0.282]	-0.654 [-0.713, -0.592]	0.318 [0.267, 0.369]
BW+	train	A	S	438	-0.180 [-0.221, -0.137]	-0.244 [-0.297, -0.196]	0.064 [0.030, 0.101]
RF−	train	A	P	720	-0.108 [-0.147, -0.067]	-0.362 [-0.410, -0.312]	0.254 [0.210, 0.300]
RF−	train	A	S	474	-0.097 [-0.133, -0.059]	-0.160 [-0.196, -0.122]	0.063 [0.032, 0.093]
RF+	train	A	P	720	-0.199 [-0.242, -0.154]	-0.362 [-0.410, -0.312]	0.164 [0.125, 0.204]
RF+	train	A	S	474	-0.129 [-0.167, -0.091]	-0.165 [-0.205, -0.124]	0.036 [0.011, 0.063]

Table 26: Llama framing-minus-neutral comparisons under both scoring policies. P/S denote first-line/strict whole-response scoring; n is each policy's valid clean/framed/neutral intersection. Δ_f and Δ_n are paired framed/neutral-minus-clean changes, and their contrast compares the same questions without outcome filtering. Neutral suffixes match unfavorable native-chat input length only; favorable rows reuse that neutral control without length matching. Brackets give pointwise 95% benchmark-stratified question-cluster intervals; strict intersections may exclude benchmarks.

G Common-Reference Geometry

Within each judge, a scalar score predictor and a high–low state contrast define a clean-score reference of rank at most two. This section connects their predictive accuracy and direction alignment to source-matched and within-cue displacement geometry. The larger logit-augmented reference is evaluated separately for Llama in Appendix I.

G.1 Clean-Score Prediction and Reference Directions

The scalar regression and high–low contrast describe different aspects of clean-score variation. Held-out regression metrics measure how well a linear function of the state predicts the judge's rating. The high–low contrast compares the mean states of answers receiving high and low clean ratings. Their joint span defines the reference separately in each model's raw activation coordinates.
Each contributing benchmark supplies equal numbers of high- and low-score examples, while its contribution to the pooled contrast depends on available support. The support tables give those counts and score ranges. Llama uses the same high–low contrast and scalar regression in the common and logit-augmented references, so these estimates and their supporting populations are reported here once.

(a) Held-out scalar-score prediction

			Llama			Qwen		
Blocks L/Q / split	N	**Rank**	**L MAE**	**L ρ**	**L R^2**	**Q MAE**	**Q ρ**	**Q R^2**
15/17/dev	240	2	0.430	0.769	0.652	0.850	0.356	-0.067
15/17/test	240	2	0.404	0.791	0.702	0.807	0.388	-0.105
23/26/dev	240	2	0.369	0.787	0.728	0.259	0.681	0.887
23/26/test	240	2	0.345	0.831	0.787	0.318	0.659	0.729
31/35/dev	240	2	0.386	0.769	0.708	0.188	0.707	0.938
31/35/test	240	2	0.340	0.826	0.784	0.245	0.695	0.834

(b) High–low reference populations

	Llama			Qwen		
Benchmark	**L t_L/t_H**	**L eligible**	**L used**	**Q t_L/t_H**	**Q eligible**	**Q used**
All	6.000/8.000	187/533	89/89	9.000/10.000	187/531	187/187
ARC-Challenge	6.000/8.000	10/230	10/10	9.000/10.000	56/184	56/56
CommonsenseQA	6.000/8.000	8/232	8/8	9.000/10.000	113/126	113/113
GSM8K	6.000/8.000	169/71	71/71	9.000/10.000	18/221	18/18

Table 27: Clean-score reference prediction and training support. L/Q denote Llama/Qwen. The common reference spans the scalar ridge coefficient and high–low contrast, of rank at most two; logit coefficients do not enter. Prediction targets the judge's own extractable rating. Eligible and used counts give low/high populations. Global 25th/75th score quantiles include ties; sampling balances low and high counts within each benchmark, not counts across benchmarks. Thresholds and question identities are shared across each model's three observed blocks.

Clean-score prediction becomes more accurate at later sampled blocks, particularly in Qwen, whose test R^2 rises from -0.105 at block 17 to 0.729 at block 26 and 0.834 at block 35. Llama's corresponding estimates are 0.702, 0.787, and 0.784. The reference therefore measures a stronger linear association with the emitted score in the later comparisons. The mean-direction energy table asks a separate question: how much of a framing-derived orientation falls within those measured scoring coordinates.

	Llama				Qwen			
Blocks L/Q / cue	**L N**	**L energy**	**L norms**	**L support**	**Q N**	**Q energy**	**Q norms**	**Q support**
15/17/BW +	8	0.9	0.096/0.995	No	61	0.5	0.073/0.997	Yes
15/17/BW −	715	42.0	0.648/0.762	Yes	700	2.2	0.148/0.989	Yes
15/17/RF +	6	14.5	0.381/0.925	No	25	0.7	0.082/0.997	No
15/17/RF −	71	1.4	0.116/0.993	Yes	36	0.0	0.018/1.000	Yes
23/26/BW +	8	0.1	0.026/1.000	No	61	27.8	0.528/0.849	Yes
23/26/BW −	715	76.2	0.873/0.488	Yes	700	69.8	0.835/0.550	Yes
23/26/RF +	6	17.9	0.423/0.906	No	25	51.3	0.716/0.698	No
23/26/RF −	71	15.0	0.387/0.922	Yes	36	45.6	0.675/0.737	Yes
31/35/BW +	8	0.5	0.072/0.997	No	61	36.4	0.603/0.798	Yes
31/35/BW −	715	77.4	0.880/0.475	Yes	700	83.3	0.913/0.408	Yes
31/35/RF +	6	17.1	0.414/0.910	No	25	48.2	0.694/0.720	No
31/35/RF −	71	12.9	0.359/0.933	Yes	36	53.3	0.730/0.683	Yes

Table 28: Effective-training mean directions in the common score reference. Energy is percent squared norm of the parallel component of a unit full mean direction, not covariance variance. Norms list $\|v_\parallel\|/\|v_\perp\|$ without renormalizing the components; their vectors sum to full. N is effective-training support and Support denotes $N \geq 30$. Low-support estimates remain descriptive. L/Q denote Llama/Qwen; BW/RF denote Bandwagon/Refinement.

For Bandwagon-negative, the captured mean-direction energy increases from 42.0% to 77.4% across the sampled Llama blocks and from 2.2% to 83.3% across the Qwen blocks. Refinement-negative has lower captured energy at the same depths. Signed cosines identify the orientations contributing to these differences: negative alignment with the high–low direction is consistent with its orientation toward higher clean ratings. Regression accuracy and coefficient-vector cosine remain distinct measurements, since accurate prediction does not require a large Euclidean cosine with every score-changing displacement.

Judge	Block	Cue	Clean ridge	Clean high–low	Neutral BW	Neutral RF	Matched frame	Matched content
L	15	BW +	0.024	-0.089	0.497	0.397	–	–
Q	17	BW +	-0.013	-0.073	0.589	0.648	–	–
L	15	BW −	-0.045	-0.646	0.485	0.394	1.000	0.517
Q	17	BW −	-0.023	-0.148	0.558	0.636	0.997	0.120
L	15	RF +	0.029	0.380	-0.233	0.315	–	–
Q	17	RF +	-0.022	-0.082	0.651	0.774	–	–
L	15	RF −	-0.006	-0.116	0.273	0.581	0.966	0.095
Q	17	RF −	-0.009	-0.017	0.634	0.746	0.953	0.037
L	23	BW +	0.012	0.024	0.325	0.268	–	–
Q	26	BW +	0.091	0.526	-0.092	-0.060	–	–
L	23	BW −	-0.055	-0.873	0.515	0.533	1.000	0.838
Q	26	BW −	-0.100	-0.835	0.563	0.611	0.996	0.807
L	23	RF +	0.025	0.423	-0.577	-0.082	–	–
Q	26	RF +	0.115	0.715	-0.175	-0.192	–	–
L	23	RF −	-0.036	-0.387	0.198	0.526	0.950	0.440
Q	26	RF −	-0.099	-0.675	0.653	0.683	0.802	0.315
L	31	BW +	0.031	0.067	0.247	0.237	–	–
Q	35	BW +	0.109	0.601	-0.242	-0.258	–	–
L	31	BW −	-0.063	-0.880	0.565	0.592	1.000	0.879
Q	35	BW −	-0.114	-0.912	0.576	0.632	0.997	0.893
L	31	RF +	0.032	0.414	-0.527	0.085	–	–
Q	35	RF +	0.113	0.692	-0.144	-0.184	–	–
L	31	RF −	-0.023	-0.359	0.213	0.650	0.948	0.450
Q	35	RF −	-0.121	-0.728	0.684	0.698	0.876	0.412

Table 29: Signed alignment between framing and comparison directions. Entries are cosines of training-fitted directions; adjacent L/Q rows compare Llama and Qwen at corresponding sampled depths. Clean-score directions point toward higher ratings and signs are unchanged. Matched frame/content use the row's cue. A dash denotes an unavailable matched direction, not zero.

G.2 Source Matching and Representation Geometry

Matching compares framing and final-answer errors at the same observed score outcome: the two arms have identical benchmark, clean-rating, and score-decrease distributions. The matching sets are constructed within each model and split because the model's ratings determine the eligible strata. Counts distinguish each arm's eligible population, selected matches, and same-question matches. Bandwagon provides the supported matched comparison in both models; the smaller Refinement populations provide descriptive comparisons under the same rule (Appendix E.3).

Mean-direction alignment measures similarity between the orientations of the two input sources, while centered spectra describe variation around each arm's mean. The decomposition adds a third measurement: how much mean-direction or displacement energy lies in the clean-score reference. A residual spectrum divides by the residual's own centered energy; its energy-retention fraction instead compares the remainder with the full displacements. Both quantities are needed to distinguish a concentrated remainder from a large one. Spectra with fewer than two pairs or zero centered energy are undefined.

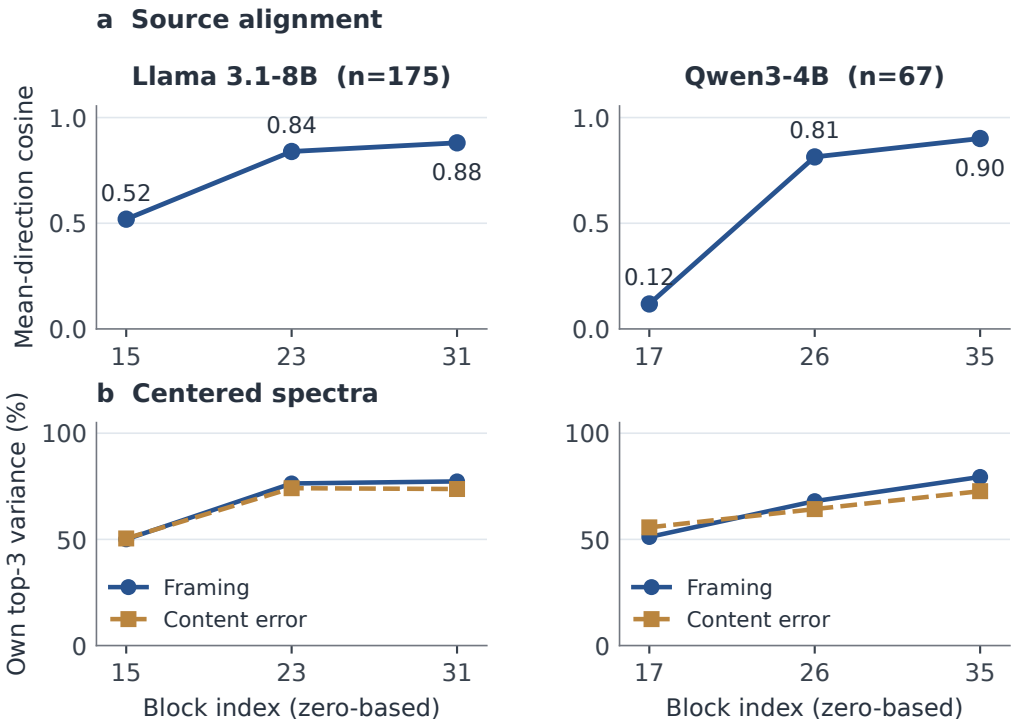


Figure 4: Source-matched Bandwagon-negative and content-error displacements. Columns give the judge and test-match count; horizontal axes give decoder blocks. Top: cosine between the two population mean displacements. Bottom: each source's own centered top-three variance fraction (framing: blue circles; content errors: amber squares). Framing and content test populations match on benchmark, clean score and observed score decrease. These are descriptive summaries without confidence intervals.

Figure 4 shows mean alignment and centered concentration across depth for the matched Bandwagon populations. The accompanying table connects each geometric comparison to its matching support: 175 held-out matches in Llama and 67 in Qwen. Mean cosine grows from 0.519 to 0.880 across the sampled Llama blocks and from 0.118 to 0.901 across Qwen blocks. Both framing and final-answer errors also become more concentrated, showing that these depth trends occur for distinct sources of matched score decreases.

(a) Eligible pools and matching support

		Llama					Qwen				
Cue	**Split**	**L N_m**	**L same q**	**L frame**	**L content**	**L primary**	**Q N_m**	**Q same q**	**Q frame**	**Q content**	**Q primary**
BW −	train	527	464	655	627	Yes	208	92	645	619	Yes
BW −	dev	173	144	228	220	Yes	63	25	225	216	Yes
BW −	test	175	150	213	204	Yes	67	26	210	200	Yes
RF −	train	20	6	66	627	No	2	2	31	619	No
RF −	dev	9	3	21	220	No	1	0	9	216	No
RF −	test	5	3	23	204	No	1	0	11	200	No

(b) Matched-source geometry across depth

			Llama					Qwen			
Blocks	**L/Q / cue/split**	N_L/N_Q	**L cos**	**L PC3 F**	**L PC3 C**	**L resid F**	**L resid C**	**Q cos**	**Q PC3 F**	**Q PC3 C**	**Q resid F**	**Q resid C**
15/17	BW−/train	527/208	0.517	48.1	51.1	90.3	93.2	0.112	49.5	50.4	98.2	98.9
15/17	BW−/test	175/67	0.519	50.0	50.4	90.8	93.9	0.118	51.1	55.7	97.7	99.1
15/17	RF−/train	20/2	0.108	45.2	63.0	95.2	94.9	0.021	100.0	100.0	98.3	99.1
15/17	RF−/test	5/1	0.141	86.8	85.9	98.4	94.2	0.019	NA	NA	NA	NA
23/26	BW−/train	527/208	0.838	75.5	74.0	65.9	63.9	0.819	65.2	62.4	74.6	78.4
23/26	BW−/test	175/67	0.840	76.3	74.1	65.6	63.1	0.814	68.0	64.3	73.9	77.3
23/26	RF−/train	20/2	0.481	50.8	77.4	90.3	70.1	0.432	100.0	100.0	55.8	79.0
23/26	RF−/test	5/1	0.565	91.1	93.0	96.9	82.5	0.585	NA	NA	NA	NA
31/35	BW−/train	527/208	0.879	76.0	73.8	68.1	62.0	0.906	78.4	69.2	62.7	74.3
31/35	BW−/test	175/67	0.880	77.3	73.7	68.0	61.7	0.901	79.4	72.7	61.5	72.0
31/35	RF−/train	20/2	0.477	50.5	74.6	90.0	68.8	0.523	100.0	100.0	45.8	64.2
31/35	RF−/test	5/1	0.550	90.3	94.1	92.1	81.7	0.669	NA	NA	NA	NA

Table 30: Support and geometry of exact score-matched framing and content changes. L/Q denote Llama/Qwen; F/C denote framing/final-answer errors. Exact strata are benchmark, clean score and positive score drop, within the correct-final-answer parent pool. N_m counts selected matches; same-question pairs are prioritized. Primary requires at least 80 training and 50 test matches, each covering at least two benchmarks. Cos compares population mean displacements. PC3 gives percent own centered variance explained by each arm's first three PCs; Resid gives percent original centered energy retained outside the rank-two reference. Estimates are descriptive; NA denotes undefined geometry. Refinement comparisons are below the primary-support requirement.

Neutral metadata and unmatched content alterations provide complementary context for these matched populations. Their spectra include all activation-paired inputs, preserving separate sources, splits, and templates. Neutral-Bandwagon template A has a final-block top-three fraction of 66.0% in Llama and 67.3% in Qwen. Concentrated displacement variation therefore also occurs under neutral appendages; the parallel-energy columns show how that variation relates to the clean-score reference.

		Llama					Qwen				
Blocks L/Q / split/input/T	N	**L PC1**	**L PC3**	**L PC5**	**L total**	**L cent.**	**Q PC1**	**Q PC3**	**Q PC5**	**Q total**	**Q cent.**
15/17/Test/Final-answer error	214	30.2	54.1	62.2	28.2	8.0	30.1	51.5	57.4	2.9	1.3
15/17/Test/Neutral BW/A	240	21.3	38.7	46.4	22.1	4.9	19.9	33.3	42.8	1.0	2.4
15/17/Test/Neutral BW/B	240	21.0	37.3	45.5	14.0	6.1	26.3	38.5	47.3	0.6	2.5
15/17/Test/Neutral RF/A	240	16.6	34.6	42.4	6.9	5.0	19.3	31.8	40.9	0.4	1.6
15/17/Test/Neutral RF/B	240	22.0	39.9	47.1	14.2	4.9	20.5	33.3	42.1	1.5	2.4
15/17/Train/Final-answer error	660	29.0	54.7	61.7	28.9	9.3	25.6	46.0	52.5	2.9	1.3
15/17/Train/Neutral BW/A	720	21.9	38.0	44.8	22.2	5.7	18.8	33.4	42.0	1.0	2.2
15/17/Train/Neutral RF/A	720	15.7	33.1	39.8	7.0	5.5	18.8	32.1	40.5	0.4	1.6
23/26/Test/Final-answer error	214	61.0	76.0	81.4	58.3	37.6	39.6	63.9	70.1	46.7	18.5
23/26/Test/Neutral BW/A	240	48.0	64.9	69.7	28.6	9.4	45.8	60.5	65.6	30.5	30.4
23/26/Test/Neutral BW/B	240	43.4	60.0	65.6	30.3	11.3	33.7	50.6	56.7	21.9	23.3
23/26/Test/Neutral RF/A	240	26.3	47.2	54.1	21.1	14.4	41.9	56.9	62.3	33.6	30.0
23/26/Test/Neutral RF/B	240	43.1	59.1	64.9	23.2	11.1	37.6	52.5	57.9	21.3	25.7
23/26/Train/Final-answer error	660	61.0	76.5	81.8	58.4	39.0	36.2	64.2	70.9	46.3	24.1
23/26/Train/Neutral BW/A	720	49.2	65.2	69.4	28.3	10.2	34.2	51.2	57.1	25.6	25.2
23/26/Train/Neutral RF/A	720	24.9	46.5	52.7	21.5	15.0	32.6	49.1	54.8	27.1	23.4
31/35/Test/Final-answer error	214	60.9	74.9	80.6	59.0	39.0	42.5	72.8	78.5	47.0	24.6
31/35/Test/Neutral BW/A	240	49.7	66.0	71.6	30.6	10.8	51.9	67.3	73.3	36.7	35.1
31/35/Test/Neutral BW/B	240	48.1	63.2	68.9	32.8	12.3	40.4	57.9	66.0	29.9	28.5
31/35/Test/Neutral RF/A	240	32.3	50.0	56.3	22.7	16.1	47.9	66.5	72.1	41.2	36.1
31/35/Test/Neutral RF/B	240	48.0	62.5	68.1	27.5	12.8	44.2	59.9	67.4	30.2	32.4
31/35/Train/Final-answer error	660	60.8	75.6	80.7	59.0	40.5	39.9	74.0	78.9	46.7	30.0
31/35/Train/Neutral BW/A	720	50.0	66.0	71.3	30.6	11.7	43.6	60.3	67.0	35.1	34.7
31/35/Train/Neutral RF/A	720	30.1	48.6	54.9	22.7	16.3	39.4	59.4	65.4	35.5	30.5

Table 31: Neutral-metadata and content-error displacement geometry across judges. L/Q denote Llama/Qwen. PC columns give percent own centered variance. Total/cent. give percent original total/centered displacement energy captured by the common scalar/high–low reference. N counts activation-paired inputs, not valid-score intersections. Sources, splits and templates remain separate. T denotes template; undefined spectra remain NA.

Within-cue projections ask whether the same training-fitted orientation is present in test responses that change the score and in those that leave it unchanged. Group membership follows each template's unsteered paired scores. The reported counts, projections, and cosines describe these groups along fixed training orientations; the full outcome-group geometry additionally includes opposite responses.

				Llama			Qwen		
Cue/T	**Group**	N_L	N_Q	$\bar{p}_F$ **(L)**	$\bar{p}_\parallel$ **(L)**	$\bar{p}_\perp$ **(L)**	$\bar{p}_F$ **(Q)**	$\bar{p}_\parallel$ **(Q)**	$\bar{p}_\perp$ **(Q)**
Test/BW +/A	All	240	240	1.816	-0.918	1.840	9.706	-2.374	12.901
Test/BW +/A	Effective	3	10	1.922	0.089	1.920	16.477	6.916	15.102
Test/BW +/A	Unchanged	185	172	1.795	-0.790	1.817	9.251	-2.523	12.458
Test/BW +/A	Opposite	52	58	1.881	-1.433	1.919	9.886	-3.536	13.835
Test/BW +/B	All	240	240	1.585	-0.068	1.588	11.348	-2.902	15.162
Test/BW +/B	Effective	19	10	1.703	0.837	1.681	22.262	11.868	18.835
Test/BW +/B	Unchanged	184	111	1.568	-0.005	1.569	10.950	-2.437	14.404
Test/BW +/B	Opposite	37	119	1.612	-0.845	1.635	10.803	-4.576	15.560
Test/BW −/A	All	240	240	5.790	5.059	2.817	50.290	42.280	27.234
Test/BW −/A	Effective	239	234	5.805	5.073	2.823	50.965	42.938	27.462
Test/BW −/A	Unchanged	1	6	2.249	1.830	1.336	23.954	16.616	18.327
Test/BW −/A	Opposite	0	0	—	—	—	—	—	—
Test/BW −/B	All	240	240	5.146	4.266	2.916	57.563	47.429	32.641
Test/BW −/B	Effective	239	238	5.159	4.278	2.919	57.930	47.806	32.735
Test/BW −/B	Unchanged	1	2	2.111	1.261	2.070	13.907	2.524	21.466
Test/BW −/B	Opposite	0	0	—	—	—	—	—	—
Test/RF +/A	All	240	240	0.829	-0.107	0.965	1.633	-0.445	2.797
Test/RF +/A	Effective	4	5	0.895	0.559	0.727	2.049	-0.177	3.118
Test/RF +/A	Unchanged	205	221	0.835	-0.075	0.956	1.915	-0.145	2.893
Test/RF +/A	Opposite	31	14	0.787	-0.407	1.058	-2.965	-5.274	1.164
Test/RF +/B	All	240	240	-0.043	-0.460	0.167	1.339	-0.176	2.099
Test/RF +/B	Effective	3	2	0.164	0.399	-0.006	1.388	-0.444	2.444
Test/RF +/B	Unchanged	202	224	-0.009	-0.404	0.179	1.641	0.111	2.238
Test/RF +/B	Opposite	35	14	-0.259	-0.855	0.113	-3.512	-4.737	-0.171
Test/RF −/A	All	240	240	0.849	-0.013	0.926	3.093	1.505	2.816
Test/RF −/A	Effective	24	14	1.070	0.399	0.993	9.205	7.744	5.390
Test/RF −/A	Unchanged	208	222	0.834	-0.038	0.921	2.703	1.110	2.649
Test/RF −/A	Opposite	8	4	0.557	-0.596	0.855	3.338	1.568	3.091
Test/RF −/B	All	240	240	0.623	0.117	0.627	1.988	0.585	2.161
Test/RF −/B	Effective	20	15	0.819	0.568	0.650	7.981	6.688	4.697
Test/RF −/B	Unchanged	208	222	0.622	0.108	0.629	1.614	0.198	2.008
Test/RF −/B	Opposite	12	3	0.321	-0.486	0.553	-0.314	-1.324	0.787

Table 32: Test displacement projections on training-fitted directions. L/Q denote Llama block 23/Qwen block 26. Directions are fitted on training questions; the common scalar/high–low span defines parallel and residual components. A/B and all/effective/unchanged/opposite groups remain separate. p projects raw activation displacements onto separately unit-normalized full, parallel and residual directions. These coordinates are not additive original-magnitude components. Counts are activation-paired group sizes. Empty groups retain zero counts and undefined entries; raw units are model-specific.

For Refinement-negative template A, Llama's effective and unchanged groups have mean parallel projections of 0.399 and -0.038, while Qwen's are 7.744 and 1.110. These are within-model comparisons in each model's activation units. Normalized alignment gives a complementary view: the effective and unchanged groups have similar residual mean cosines, 0.653 versus 0.636 in Llama and 0.426 versus 0.392 in Qwen. A residual orientation can thus be present even when the emitted score remains unchanged, while the parallel projections distinguish these observed response groups more clearly.

		Llama			Qwen		
Cue/T	**Group**	$\bar{c}_F$ **(L)**	$\bar{c}_\parallel$ **(L)**	$\bar{c}_\perp$ **(L)**	$\bar{c}_F$ **(Q)**	$\bar{c}_\parallel$ **(Q)**	$\bar{c}_\perp$ **(Q)**
Test/BW +/A	All	0.670	-0.317	0.679	0.501	-0.127	0.669
Test/BW +/A	Effective	0.688	0.060	0.687	0.788	0.332	0.721
Test/BW +/A	Unchanged	0.687	-0.287	0.695	0.484	-0.136	0.654
Test/BW +/A	Opposite	0.611	-0.444	0.623	0.504	-0.179	0.704
Test/BW +/B	All	0.586	-0.026	0.587	0.468	-0.129	0.632
Test/BW +/B	Effective	0.550	0.260	0.543	0.780	0.415	0.661
Test/BW +/B	Unchanged	0.594	-0.003	0.594	0.463	-0.108	0.612
Test/BW +/B	Opposite	0.568	-0.289	0.576	0.447	-0.195	0.648
Test/BW −/A	All	0.908	0.786	0.456	0.915	0.766	0.501
Test/BW −/A	Effective	0.910	0.787	0.456	0.921	0.774	0.499
Test/BW −/A	Unchanged	0.554	0.451	0.329	0.683	0.437	0.579
Test/BW −/A	Opposite	—	—	—	—	—	—
Test/BW −/B	All	0.870	0.706	0.520	0.906	0.743	0.519
Test/BW −/B	Effective	0.871	0.707	0.519	0.910	0.748	0.518
Test/BW −/B	Unchanged	0.639	0.382	0.626	0.424	0.076	0.656
Test/BW −/B	Opposite	—	—	—	—	—	—
Test/RF +/A	All	0.518	-0.069	0.604	0.267	-0.060	0.444
Test/RF +/A	Effective	0.550	0.368	0.435	0.306	-0.021	0.459
Test/RF +/A	Unchanged	0.526	-0.053	0.605	0.291	-0.039	0.456
Test/RF +/A	Opposite	0.463	-0.233	0.620	-0.124	-0.411	0.245
Test/RF +/B	All	0.006	-0.278	0.136	0.235	-0.012	0.350
Test/RF +/B	Effective	0.127	0.281	0.009	0.220	-0.041	0.357
Test/RF +/B	Unchanged	0.025	-0.255	0.146	0.258	0.007	0.362
Test/RF +/B	Opposite	-0.114	-0.460	0.089	-0.123	-0.321	0.154
Test/RF −/A	All	0.584	-0.003	0.635	0.408	0.174	0.394
Test/RF −/A	Effective	0.696	0.243	0.653	0.676	0.536	0.426
Test/RF −/A	Unchanged	0.579	-0.017	0.636	0.391	0.151	0.392
Test/RF −/A	Opposite	0.365	-0.383	0.557	0.391	0.151	0.391
Test/RF −/B	All	0.410	0.079	0.412	0.267	0.059	0.308
Test/RF −/B	Effective	0.519	0.352	0.415	0.518	0.404	0.333
Test/RF −/B	Unchanged	0.411	0.075	0.415	0.254	0.039	0.309
Test/RF −/B	Opposite	0.210	-0.291	0.350	-0.020	-0.184	0.142

Table 33: Normalized test displacement alignment on training-fitted directions. L/Q denote Llama block 23/Qwen block 26. Directions are fitted on training questions; the common scalar/high–low span defines parallel and residual components. A/B and all/effective/unchanged/opposite groups remain separate. c is mean cosine to each direction, omitting zero-norm displacements. Group support is given in Table 32; both tables use the same rows and fixed orientations. Empty groups retain undefined entries. Entries are descriptive point estimates.

H Common-Reference Interventions

H.1 Forward Effects, Ranking, and Scoring Policy

The forward comparison tests whether a framing-derived direction changes clean ratings differently from score, neutral-metadata, and matched-content directions. The tables include the full direction, both reference components, all five random draws, and each available matched-source direction. Every unit direction receives its own development selection over the shared signed grid. Raw components keep their projected magnitudes and the full direction's coefficient; their development feasibility therefore refers to the full direction, while their test effects and rank correlations measure the actual component intervention. Figure 5 visualizes the component and clean-score comparisons, including uncertainty and test-fidelity status.

Positive effects denote mean decreases in clean ratings. For each scoring policy, the conservative endpoint uses a fixed set of initially valid clean references and assigns -9 when an intervention produces an invalid rating. The valid-pair mean and its coverage accompany this endpoint (Appendix E.5). Primary and strict-format evaluations use the same fitted direction, strength, and completion.

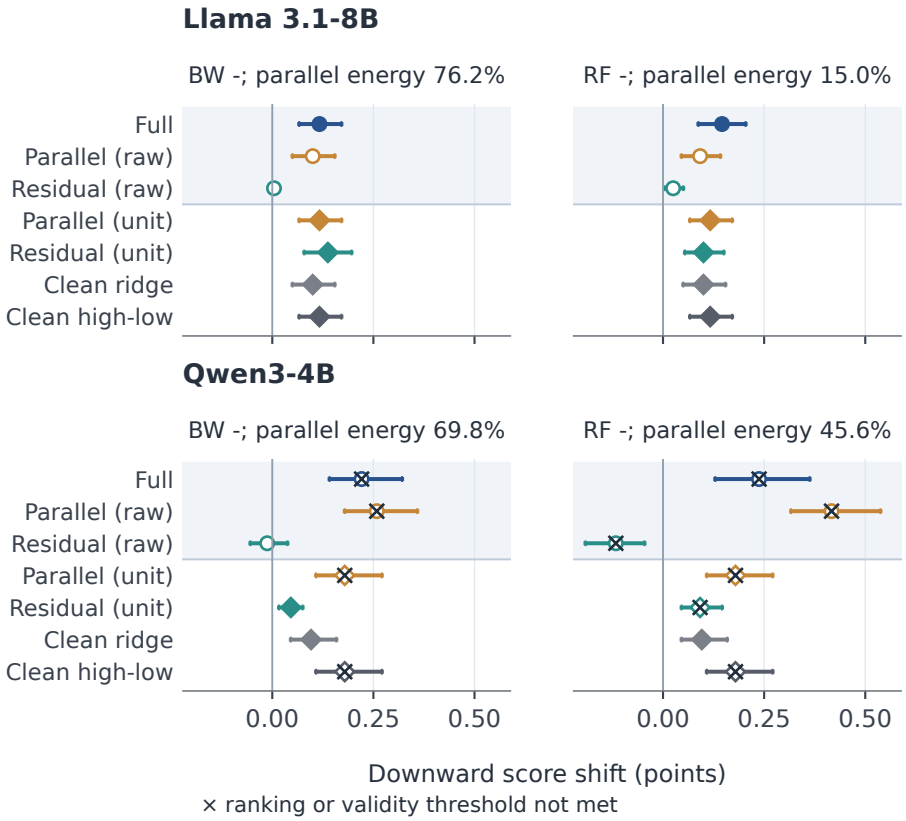


Figure 5: Clean-input score control with the within-judge rank-two scalar/high–low reference. Shading groups full with raw components, which inherit its selected coefficient without renormalization (open circles); unit components and clean references use independently development-selected coefficients. Energy labels report the parallel component's squared norm relative to the unit full direction. Points show primary mean decreases; bars are 95% benchmark-stratified question-bootstrap intervals on fixed valid-clean support (later invalid outputs: -9). Crosses mark failure of the test ranking or validity thresholds ($\rho \geq 0.85$ and validity ≥ 0.93).

Direction	Llama L α N_P/N_S	Llama L shift [95% CI]	Llama L ρ D/T	Qwen Q α N_P/N_S	Qwen Q shift [95% CI]	Qwen Q ρ D/T
Full	0.350 240/162	P: 0.117 [0.067, 0.171] S: -0.086 [-0.261, 0.032]	0.919 Y/Y	11.129 240/240	P: 0.221 [0.142, 0.321] S: 0.221 [0.142, 0.321]	0.804 Y/N
Parallel (raw)	0.350 240/162	P: 0.100 [0.050, 0.154] S: -0.142 [-0.351, 0.025]	0.924 Y/Y	11.129 240/240	P: 0.258 [0.179, 0.358] S: 0.258 [0.179, 0.358]	0.737 Y/N
Residual (raw)	0.350 240/162	P: 0.004 [0.000, 0.013] S: 0.006 [0.000, 0.019]	0.995 Y/Y	11.129 240/240	P: -0.013 [-0.054, 0.037] S: -0.013 [-0.054, 0.037]	0.867 Y/Y
Parallel (unit)	0.350 240/162	P: 0.117 [0.067, 0.171] S: -0.142 [-0.351, 0.025]	0.919 Y/Y	5.565 240/240	P: 0.179 [0.108, 0.271] S: 0.179 [0.108, 0.271]	0.806 Y/N
Residual (unit)	1.401 240/162	P: 0.138 [0.079, 0.196] S: 0.000 [-0.147, 0.082]	0.895 Y/Y	-5.565 240/240	P: 0.046 [0.017, 0.075] S: 0.046 [0.017, 0.075]	0.898 Y/Y
Clean ridge	-1.401 240/162	P: 0.100 [0.050, 0.154] S: -0.031 [-0.182, 0.043]	0.919 Y/Y	-11.129 240/240	P: 0.096 [0.046, 0.158] S: 0.096 [0.046, 0.158]	0.856 Y/Y
Clean high-low	-0.350 240/162	P: 0.117 [0.067, 0.171] S: -0.142 [-0.351, 0.025]	0.919 Y/Y	-5.565 240/240	P: 0.179 [0.108, 0.271] S: 0.179 [0.108, 0.271]	0.806 Y/N
Neutral	0.700 240/162	P: 0.117 [0.067, 0.171] S: -0.142 [-0.351, 0.025]	0.919 Y/Y	22.259 240/240	P: 0.221 [0.146, 0.321] S: 0.221 [0.146, 0.321]	0.769 Y/N
Random 101	2.802 240/162	P: 0.013 [-0.013, 0.037] S: -0.086 [-0.261, 0.038]	0.968 Y/Y	-22.259 240/240	P: 0.125 [0.067, 0.200] S: 0.125 [0.067, 0.200]	0.844 Y/N
Random 102	2.802 240/162	P: 0.008 [-0.046, 0.058] S: 0.074 [0.012, 0.135]	0.863 Y/Y	22.259 240/240	P: 0.062 [0.021, 0.108] S: 0.062 [0.021, 0.108]	0.885 Y/Y
Random 103	-2.802 240/162	P: 0.046 [0.021, 0.075] S: 0.056 [0.024, 0.093]	0.957 Y/Y	-22.259 240/240	P: 0.083 [0.025, 0.154] S: 0.083 [0.025, 0.154]	0.903 Y/Y
Random 104	2.802 240/162	P: 0.042 [0.000, 0.083] S: -0.074 [-0.220, -0.006]	0.949 Y/Y	-11.129 240/240	P: 0.021 [0.004, 0.042] S: 0.021 [0.004, 0.042]	0.969 Y/Y
Random 105	-2.802 240/162	P: 0.058 [0.021, 0.100] S: -0.099 [-0.269, 0.019]	0.958 Y/Y	11.129 240/240	P: 0.075 [0.033, 0.125] S: 0.075 [0.033, 0.125]	0.861 Y/Y
Matched frame	0.350 240/162	P: 0.117 [0.067, 0.171] S: -0.086 [-0.261, 0.032]	0.919 Y/Y	11.129 240/240	P: 0.208 [0.129, 0.308] S: 0.208 [0.129, 0.308]	0.823 Y/N
Matched content	0.350 240/162	P: 0.096 [0.050, 0.146] S: -0.093 [-0.267, 0.025]	0.935 Y/Y	22.259 240/240	P: 0.362 [0.233, 0.500] S: 0.362 [0.233, 0.500]	0.756 Y/N

Table 34: Bandwagon-negative clean-forward comparisons across judges. L/Q denote Llama/Qwen; P/S are first-line/strict scoring. Positive shift means a lower rating. Each policy uses its fixed valid-clean support N_P/N_S, with -9 for a later invalid output. The same directions, coefficients and completions underlie both policies. Raw components keep full's coefficient and projected magnitude; unit directions have separate development selection. D/T give development feasibility/test fidelity. For raw components D inherits full's feasibility; T reports actual component fidelity. Infeasible searches use zero intervention; matched directions below support are unavailable. Intervals are 95% benchmark-stratified joint-question bootstrap intervals (2,000 resamples). All evaluated Qwen methods have 240/240/240/0/0 inputs/first-line/whole-response/suffix/cap counts, so their valid-pair and fixed-support estimates coincide.

The original-magnitude Bandwagon components have different score effects. At the full direction's coefficient, the parallel component lowers ratings by 0.100 points in Llama and 0.258 in Qwen, while the residual effects are 0.004 and -0.013. Giving the residual its own unit normalization and development selection changes the comparison: its effects become 0.138 and 0.046. These results separate the score effect retained at the original component magnitude from the effect obtainable along a rescaled orientation.

Direction	Llama			Qwen		
	L α N_P/N_S	L shift [95% CI]	L ρ D/T	Q α N_P/N_S	Q shift [95% CI]	Q ρ D/T
Full	0.700 240/162	P: 0.146 [0.087, 0.204] S: 0.012 [-0.135, 0.099]	0.886 Y/Y	22.259 240/240	P: 0.237 [0.129, 0.362] S: 0.237 [0.129, 0.362]	0.833 Y/N
Parallel (raw)	0.700 240/162	P: 0.092 [0.046, 0.142] S: -0.086 [-0.261, 0.032]	0.930 Y/Y	22.259 240/240	P: 0.417 [0.317, 0.537] S: 0.417 [0.317, 0.537]	0.653 Y/N
Residual (raw)	0.700 240/162	P: 0.025 [0.004, 0.050] S: -0.031 [-0.182, 0.043]	0.974 Y/Y	22.259 240/240	P: -0.117 [-0.192, -0.046] S: -0.117 [-0.192, -0.046]	0.628 Y/N
Parallel (unit)	0.350 240/162	P: 0.117 [0.067, 0.171] S: -0.142 [-0.351, 0.025]	0.919 Y/Y	5.565 240/240	P: 0.179 [0.108, 0.271] S: 0.179 [0.108, 0.271]	0.806 Y/N
Residual (unit)	1.401 240/162	P: 0.100 [0.054, 0.150] S: 0.019 [-0.129, 0.106]	0.911 Y/Y	-5.565 240/240	P: 0.092 [0.046, 0.146] S: 0.092 [0.046, 0.146]	0.832 Y/N
Clean ridge	-1.401 240/162	P: 0.100 [0.050, 0.154] S: -0.031 [-0.182, 0.043]	0.919 Y/Y	-11.129 240/240	P: 0.096 [0.046, 0.158] S: 0.096 [0.046, 0.158]	0.856 Y/Y
Clean high-low	-0.350 240/162	P: 0.117 [0.067, 0.171] S: -0.142 [-0.351, 0.025]	0.919 Y/Y	-5.565 240/240	P: 0.179 [0.108, 0.271] S: 0.179 [0.108, 0.271]	0.806 Y/N
Neutral	0.350 240/162	P: 0.058 [0.021, 0.100] S: -0.099 [-0.272, 0.019]	0.958 Y/Y	11.129 240/240	P: 0.158 [0.092, 0.246] S: 0.158 [0.092, 0.246]	0.834 Y/N
Random 101	2.802 240/162	P: 0.013 [-0.013, 0.037] S: -0.086 [-0.261, 0.038]	0.968 Y/Y	-22.259 240/240	P: 0.125 [0.067, 0.200] S: 0.125 [0.067, 0.200]	0.844 Y/N
Random 102	2.802 240/162	P: 0.008 [-0.046, 0.058] S: 0.074 [0.012, 0.135]	0.863 Y/Y	22.259 240/240	P: 0.062 [0.021, 0.108] S: 0.062 [0.021, 0.108]	0.885 Y/Y
Random 103	-2.802 240/162	P: 0.046 [0.021, 0.075] S: 0.056 [0.024, 0.093]	0.957 Y/Y	-22.259 240/240	P: 0.083 [0.025, 0.154] S: 0.083 [0.025, 0.154]	0.903 Y/Y
Random 104	2.802 240/162	P: 0.042 [0.000, 0.083] S: -0.074 [-0.220, -0.006]	0.949 Y/Y	-11.129 240/240	P: 0.021 [0.004, 0.042] S: 0.021 [0.004, 0.042]	0.969 Y/Y
Random 105	-2.802 240/162	P: 0.058 [0.021, 0.100] S: -0.099 [-0.269, 0.019]	0.958 Y/Y	11.129 240/240	P: 0.075 [0.033, 0.125] S: 0.075 [0.033, 0.125]	0.861 Y/Y
Matched frame	– –	P: Below support S: –	– –	– –	P: Below support S: –	– –
Matched content	– –	P: Below support S: –	– –	– –	P: Below support S: –	– –

Table 35: Refinement-negative clean-forward comparisons across judges. L/Q denote Llama/Qwen; P/S are first-line/strict scoring. Positive shift means a lower rating. Each policy uses its fixed valid-clean support N_P/N_S, with -9 for a later invalid output. The same directions, coefficients and completions underlie both policies. Raw components keep full's coefficient and projected magnitude; unit directions have separate development selection. D/T give development feasibility/test fidelity. For raw components D inherits full's feasibility; T reports actual component fidelity. Infeasible searches use zero intervention; matched directions below support are unavailable. Intervals are 95% benchmark-stratified joint-question bootstrap intervals (2,000 resamples). All evaluated Qwen methods have 240/240/240/0/0 inputs/first-line/whole-response/suffix/cap counts, so their valid-pair and fixed-support estimates coincide.

Refinement exhibits the same need to read mean effects together with ranking. Its full direction lowers Llama scores by 0.146 points with test correlation 0.886. Qwen's full effect is 0.237 with correlation 0.833, below the 0.85 criterion; its raw parallel component produces a larger decrease but a lower correlation. Development feasibility therefore does not guarantee the same fidelity on held-out questions. The policy and validity diagnostics in Appendix H.3 show which emitted ratings contribute to each estimate.

H.2 Paired Comparisons with Control Directions

Full-minus-control intervals use paired differences on the same questions. The comparisons include both clean-score orientations, neutral-metadata directions, matched-content directions, and every random draw. These contrasts measure differences in score effects at each direction's development-selected setting; the corresponding rank correlations and validity counts appear in Appendix H.1.

Control	L P contrast [95% CI]	L S contrast [95% CI]	Q P=S contrast [95% CI]
Parallel (raw)	0.017 [0.000, 0.042]	0.056 [0.000, 0.168]	-0.038 [-0.075, 0.000]
Residual (raw)	0.113 [0.058, 0.167]	-0.093 [-0.267, 0.025]	0.233 [0.158, 0.321]
Parallel (unit)	0.000 [0.000, 0.000]	0.056 [0.000, 0.168]	0.042 [0.008, 0.083]
Residual (unit)	-0.021 [-0.038, -0.004]	-0.086 [-0.206, -0.012]	0.175 [0.096, 0.271]
Clean ridge	0.017 [0.000, 0.042]	-0.056 [-0.169, 0.000]	0.125 [0.062, 0.200]
Clean high-low	0.000 [0.000, 0.000]	0.056 [0.000, 0.168]	0.042 [0.008, 0.083]
Neutral	0.000 [0.000, 0.000]	0.056 [0.000, 0.168]	0.000 [-0.038, 0.042]
Random 101	0.104 [0.054, 0.158]	0.000 [-0.166, 0.168]	0.096 [0.042, 0.154]
Random 102	0.108 [0.042, 0.179]	-0.160 [-0.340, -0.030]	0.158 [0.087, 0.242]
Random 103	0.071 [0.021, 0.129]	-0.142 [-0.317, -0.019]	0.138 [0.083, 0.204]
Random 104	0.075 [0.037, 0.121]	-0.012 [-0.139, 0.067]	0.200 [0.121, 0.300]
Random 105	0.058 [0.021, 0.104]	0.012 [-0.151, 0.176]	0.146 [0.075, 0.229]
Matched frame	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.012 [0.000, 0.029]
Matched content	0.021 [0.000, 0.050]	0.006 [0.000, 0.019]	-0.142 [-0.242, -0.054]

Table 36: Bandwagon-negative paired full-minus-control score-shift differences. L/Q denote Llama/Qwen; P/S denote first-line/strict scoring. Positive means full decreases ratings more than the control. Intervals use joint same-question bootstrap contrasts of the conservative fixed-reference endpoint, not differences between separate interval endpoints. Policy-specific reference counts are L P: 240; L S: 162; Q P=S: 240. All available controls and all five random draws are reported. Each direction uses its own development-selected setting; the corresponding ranking and validity results accompany the forward comparison.

Bandwagon's nonrandom controls contextualize its advantage over random orientations. Under primary scoring, Llama's full and neutral directions have equal mean effects, while its full-minus-matched-content difference is 0.021 points. Qwen's matched-content direction produces a larger decrease than full, with a paired full-minus-content contrast of -0.142 points (95% CI $[-0.242, -0.054]$). Both of those Qwen settings fall below the test ranking threshold. The control comparison therefore concerns both the magnitude of score movement and the ordering retained at the selected setting.

Control	L P contrast [95% CI]	L S contrast [95% CI]	Q P=S contrast [95% CI]
Parallel (raw)	0.054 [0.021, 0.092]	0.099 [0.019, 0.221]	-0.179 [-0.246, -0.112]
Residual (raw)	0.121 [0.067, 0.179]	0.043 [0.013, 0.075]	0.354 [0.233, 0.496]
Parallel (unit)	0.029 [0.008, 0.050]	0.154 [0.031, 0.323]	0.058 [-0.025, 0.146]
Residual (unit)	0.046 [0.008, 0.092]	-0.006 [-0.019, 0.000]	0.146 [0.042, 0.263]
Clean ridge	0.046 [0.017, 0.079]	0.043 [0.013, 0.075]	0.142 [0.042, 0.250]
Clean high-low	0.029 [0.008, 0.050]	0.154 [0.031, 0.323]	0.058 [-0.025, 0.146]
Neutral	0.088 [0.042, 0.138]	0.111 [0.031, 0.237]	0.079 [0.000, 0.167]
Random 101	0.133 [0.075, 0.192]	0.099 [-0.091, 0.307]	0.112 [0.021, 0.212]
Random 102	0.138 [0.075, 0.204]	-0.062 [-0.209, 0.031]	0.175 [0.071, 0.292]
Random 103	0.100 [0.050, 0.154]	-0.043 [-0.189, 0.025]	0.154 [0.071, 0.250]
Random 104	0.104 [0.062, 0.154]	0.086 [0.048, 0.132]	0.217 [0.112, 0.338]
Random 105	0.088 [0.042, 0.138]	0.111 [0.031, 0.238]	0.162 [0.058, 0.279]

Table 37: Refinement-negative paired full-minus-control score-shift differences. L/Q denote Llama/Qwen; P/S denote first-line/strict scoring. Positive means full decreases ratings more than the control. Intervals use joint same-question bootstrap contrasts of the conservative fixed-reference endpoint, not differences between separate interval endpoints. Policy-specific reference counts are L P: 240; L S: 162; Q P=S: 240. All available controls and all five random draws are reported. Each direction uses its own development-selected setting; the corresponding ranking and validity results accompany the forward comparison.

For Refinement, the primary full-minus-neutral contrast is 0.088 points in Llama (95% CI $[0.042, 0.138]$) and 0.079 in Qwen ($[0.000, 0.167]$). The five random draws remain separate comparisons under the same development search. The paired intervals use the joint question-level differences directly, preserving the dependence between each full and control response.

H.3 Output Format after Intervention

The Llama diagnostics place complete-valid-pair effects beside output-format counts for every selected method, including raw components and zero-strength configurations. The fixed-reference conservative endpoint includes a penalty for a later invalid score; the valid-pair mean instead describes the responses for which both ratings can be extracted. Displaying their coverage makes the two readings comparable without treating them as the same population. Qwen's selected outputs satisfy both rating policies, whose test estimates coincide in the corresponding effect tables.

	First-line		Strict		Output coverage	
Direction	**P valid shift [95% CI]**	**P N_p/N_r**	**S valid shift [95% CI]**	**S N_p/N_r**	**Inputs/ first**	**Whole/ suffix/cap**
Full	0.117 [0.067, 0.171]	240/240	0.025 [0.006, 0.050]	160/162	240/240	161/79/79
Parallel (raw)	0.100 [0.050, 0.154]	240/240	0.025 [0.006, 0.050]	159/162	240/240	159/81/81
Residual (raw)	0.004 [0.000, 0.013]	240/240	0.006 [0.000, 0.019]	162/162	240/240	162/78/78
Parallel (unit)	0.117 [0.067, 0.171]	240/240	0.025 [0.006, 0.050]	159/162	240/240	160/80/80
Residual (unit)	0.138 [0.079, 0.196]	240/240	0.056 [0.024, 0.094]	161/162	240/240	162/78/78
Clean ridge	0.100 [0.050, 0.154]	240/240	0.025 [0.006, 0.050]	161/162	240/240	162/78/78
Clean high-low	0.117 [0.067, 0.171]	240/240	0.025 [0.006, 0.050]	159/162	240/240	160/80/80
Neutral	0.117 [0.067, 0.171]	240/240	0.025 [0.006, 0.050]	159/162	240/240	160/80/80
Random 101	0.013 [-0.013, 0.037]	240/240	0.025 [0.000, 0.056]	160/162	240/240	160/80/80
Random 102	0.008 [-0.046, 0.058]	240/240	0.074 [0.012, 0.135]	162/162	240/240	162/78/78
Random 103	0.046 [0.021, 0.075]	240/240	0.056 [0.024, 0.093]	162/162	240/240	162/78/78
Random 104	0.042 [0.000, 0.083]	240/240	-0.019 [-0.043, 0.000]	161/162	240/240	161/79/79
Random 105	0.058 [0.021, 0.100]	240/240	0.013 [0.000, 0.031]	160/162	240/240	160/80/80
Matched frame	0.117 [0.067, 0.171]	240/240	0.025 [0.006, 0.050]	160/162	240/240	161/79/79
Matched content	0.096 [0.050, 0.146]	240/240	0.019 [0.000, 0.043]	160/162	240/240	160/80/80

Table 38: Llama Bandwagon-negative valid-pair effects and output coverage. P/S denote first-line/strict scoring. Positive shift means a lower rating. N_p/N_r counts valid clean–steered pairs/fixed originally valid clean references. Valid-pair means exclude later invalid outputs and supplement the conservative endpoint. Output counts instead use all queried clean inputs: First is first-line validity, Whole is integer-only whole-response compliance, Suffix is nonempty subsequent text, and Cap is decoder-limit termination. These counts are distinct from paired support. Raw-component output counts come from the actual component intervention. Intervals are 95% benchmark-stratified question-bootstrap intervals; strict scoring reuses primary selection.

For Llama's full Bandwagon direction, the strict valid-pair mean decrease is 0.025 points, whereas the conservative strict effect is -0.086. The former uses 160 valid clean–steered pairs out of 162 initially valid clean references. The output itself satisfies the whole-response rule on 161 of the 240 queried inputs, a different count because it does not require a valid clean reference. This distinction explains why output compliance and paired-effect coverage are recorded separately.

Direction	First-line P valid shift [95% CI]	P N_p/N_r	Strict S valid shift [95% CI]	S N_p/N_r	Output coverage Inputs/ first	Whole/ suffix/cap
Full	0.146 [0.087, 0.204]	240/240	0.068 [0.031, 0.108]	161/162	240/240	162/78/78
Parallel (raw)	0.092 [0.046, 0.142]	240/240	0.025 [0.006, 0.050]	160/162	240/240	160/80/80
Residual (raw)	0.025 [0.004, 0.050]	240/240	0.025 [0.006, 0.050]	161/162	240/240	161/79/79
Parallel (unit)	0.117 [0.067, 0.171]	240/240	0.025 [0.006, 0.050]	159/162	240/240	160/80/80
Residual (unit)	0.100 [0.054, 0.150]	240/240	0.075 [0.037, 0.118]	161/162	240/240	161/79/79
Clean ridge	0.100 [0.050, 0.154]	240/240	0.025 [0.006, 0.050]	161/162	240/240	162/78/78
Clean high-low	0.117 [0.067, 0.171]	240/240	0.025 [0.006, 0.050]	159/162	240/240	160/80/80
Neutral	0.058 [0.021, 0.100]	240/240	0.013 [0.000, 0.031]	160/162	240/240	160/80/80
Random 101	0.013 [-0.013, 0.037]	240/240	0.025 [0.000, 0.056]	160/162	240/240	160/80/80
Random 102	0.008 [-0.046, 0.058]	240/240	0.074 [0.012, 0.135]	162/162	240/240	162/78/78
Random 103	0.046 [0.021, 0.075]	240/240	0.056 [0.024, 0.093]	162/162	240/240	162/78/78
Random 104	0.042 [0.000, 0.083]	240/240	-0.019 [-0.043, 0.000]	161/162	240/240	161/79/79
Random 105	0.058 [0.021, 0.100]	240/240	0.013 [0.000, 0.031]	160/162	240/240	160/80/80
Matched frame	Below support	–	–	–	—	–/–/–
Matched content	Below support	–	–	–	—	–/–/–

Table 39: Llama Refinement-negative valid-pair effects and output coverage. P/S denote first-line/strict scoring. Positive shift means a lower rating. N_p/N_r counts valid clean–steered pairs/fixed originally valid clean references. Valid-pair means exclude later invalid outputs and supplement the conservative endpoint. Output counts instead use all queried clean inputs: First is first-line validity, Whole is integer-only whole-response compliance, Suffix is nonempty subsequent text, and Cap is decoder-limit termination. These counts are distinct from paired support. Raw-component output counts come from the actual component intervention. Intervals are 95% benchmark-stratified question-bootstrap intervals; strict scoring reuses primary selection.

Refinement's full direction retains all 240 first-line pairs and 161 of the 162 strict clean references, with valid-pair mean decreases of 0.146 and 0.068. Its independently selected unit residual has a strict valid-pair decrease of 0.075, compared with 0.025 for the original-magnitude residual. The output diagnostics keep the effects of component scaling separate from the number of usable rating pairs.

I Llama Geometry with a Logit-Augmented Score Reference

Adding a centered-logit predictor broadens the measured clean-score reference beyond the scalar predictor and high–low contrast. Their joint span has rank at most eleven, compared with at most two in Appendix G. The Llama inputs, scores, full directions, and matched pairs are the same in both analyses; the broader span changes the parallel and residual components. Response counts and matching support are therefore given once in Appendices F.1 and G.2, and the shared scalar and high–low estimates appear in Appendix G.1.

I.1 Readout Accuracy and Reference Energy

The centered-logit regression predicts the centered logits of the ten rating tokens from clean states. Development questions select regularization, and held-out questions evaluate predictive accuracy at each observed block. At block 23, clean-test R^2 is 0.991. The resulting coefficient span contributes additional empirically measured scoring coordinates to the reference. Its union rank with the scalar coefficient and high–low contrast is distinct from the logit predictor's rank and from the number of principal components describing paired displacements.

For each framing condition and block, the mean-direction energy fraction measures how much of the effective-training mean lies in this broader reference. Table 40 aligns the readout and energy measurements by depth. Logit prediction becomes more accurate in later blocks, but the two unfavorable cues allocate very different shares of their mean direction to the reference: Bandwagon rises from 42.3% to 78.2%, whereas Refinement rises from 2.1% to 13.9%. Predictive accuracy characterizes the clean-score reference; energy characterizes a particular framing direction's relationship to it.

Readout accuracy on clean questions

Split	N_s/N_l	Block 15 R_l^2	Block 23 R_l^2	Block 31 R_l^2	Rank
dev	240/240	0.976	0.992	0.999	11
test	240/240	0.975	0.991	0.999	11

Effective-training mean-direction energy

Cue	N_{eff}	Support	Block 15 (%)	Block 23 (%)	Block 31 (%)
BW +	8	No	1.5	2.0	2.7
BW −	715	Yes	42.3	76.7	78.2
RF +	6	No	15.6	18.5	18.2
RF −	71	Yes	2.1	15.6	13.9

Table 40: Logit-augmented clean-score reference in Llama. N_s/N_l counts scalar ratings and centered ten-score-token logit targets. R_l^2 evaluates the development-selected ridge predictor. Rank is the joint scalar, high–low, and logit-coefficient span, at most eleven. Energy is squared projection norm of a unit mean direction, not centered covariance variance. Support denotes $N_{\text{eff}} \geq 30$; lower-support conditions are descriptive. BW/RF: Bandwagon/Refinement.

I.2 Displacement Geometry and Frozen-Direction Projections

The outcome-group comparison relates the size and concentration of displacement to its position along a training-fitted direction. Template-A groups contain all pairs, effective responses, unchanged scores, or opposite responses, as determined by their unsteered paired scores. All-case rows include every pair with both activation endpoints. Each residual top-three fraction divides by the residual's own centered energy, whereas retained total and centered fractions divide by the corresponding full displacement energy. The two measurements distinguish how concentrated the remainder is from how much energy remains.

Each depth-wise table places these geometric measurements beside projections of the same groups. Projection coordinates use unit full, parallel, and residual directions fitted on training questions. The independently normalized component coordinates are therefore not additive original-magnitude components. Cosine means describe nonzero displacements. At block 15, Refinement− already has positive residual alignment in both effective and unchanged groups, with cosines 0.710 and 0.694. Its parallel projections differ more clearly between these groups than its residual projections.

Cue	**Group**	N_h	F_3	R_3	E_T	E_C	$\bar{p}_F$	$\bar{p}_\parallel$	$\bar{p}_\perp$	$\bar{c}_F$	$\bar{c}_\perp$
BW +	All	240	41.8	39.8	93.1	93.7	1.519	0.373	1.484	0.858	0.839
BW +	Effective	3	100.0	100.0	97.5	97.4	1.542	0.240	1.524	0.857	0.848
BW +	Unchanged	185	41.7	39.9	93.7	94.2	1.512	0.357	1.479	0.860	0.842
BW +	Opposite	52	47.2	44.9	90.7	93.1	1.543	0.440	1.500	0.850	0.827
BW −	All	240	49.0	46.0	62.8	89.0	2.235	1.464	1.689	0.910	0.691
BW −	Effective	239	48.9	45.9	62.7	89.0	2.237	1.466	1.690	0.910	0.691
BW −	Unchanged	1	NA	NA	76.1	NA	1.769	1.049	1.430	0.813	0.657
BW −	Opposite	—	—	—	—	—	—	—	—	—	—
RF +	All	240	35.4	33.6	96.1	94.5	0.534	0.082	0.546	0.621	0.638
RF +	Effective	4	100.0	100.0	94.3	97.6	0.530	0.182	0.498	0.607	0.569
RF +	Unchanged	205	35.4	33.6	96.0	94.5	0.531	0.086	0.541	0.619	0.634
RF +	Opposite	31	44.3	42.6	96.7	94.7	0.552	0.044	0.582	0.633	0.668
RF −	All	240	33.7	31.5	96.4	94.0	0.571	-0.001	0.577	0.688	0.695
RF −	Effective	24	44.1	43.5	97.5	96.3	0.587	0.066	0.584	0.713	0.710
RF −	Unchanged	208	33.8	31.6	96.4	94.1	0.570	-0.004	0.577	0.687	0.694
RF −	Opposite	8	67.2	66.8	93.0	96.9	0.541	-0.131	0.566	0.637	0.667

Table 41: Outcome-group geometry and training-direction projections at Llama block 15. Template A; each row uses the same N_h activation-paired inputs for both sets of measurements. F_3/R_3 are full/residual top-three centered-variance percentages, each using its own branch denominator. E_T/E_C are retained residual percentages of original full total/centered energy. $\bar{p}$ is mean displacement projection in raw activation units onto a unit full, parallel, or residual training-fitted orientation; these separately normalized coordinates are not additive raw components. $\bar{c}$ averages cosines over nonzero displacements. All includes invalid-score pairs; response groups require valid paired scores. NA: undefined; —: empty or unavailable, not zero. BW/RF: Bandwagon/Refinement.

At block 23, the 24 effective Refinement− displacements have mean parallel projection 0.410, compared with -0.026 for the 208 unchanged displacements. Their residual projections are 0.989 and 0.919, with residual cosines 0.650 and 0.634. This comparison locates the difference between outcome groups along the measured score reference while also showing substantial residual alignment when ratings stay unchanged. Bandwagon− has a different group composition: its effective group contains 239 of the 240 pairs, so its effective and all-pair summaries are nearly identical.

Cue	Group	N_h	F_3	R_3	E_T	E_C	$\bar{p}_F$	$\bar{p}_\parallel$	$\bar{p}_\perp$	$\bar{c}_F$	$\bar{c}_\perp$
BW +	All	240	53.1	47.1	79.8	84.0	1.816	0.077	1.823	0.670	0.673
BW +	Effective	3	100.0	100.0	93.9	87.6	1.922	0.201	1.913	0.688	0.686
BW +	Unchanged	185	50.2	46.3	83.5	86.9	1.795	0.093	1.800	0.687	0.689
BW +	Opposite	52	59.4	55.7	69.9	85.4	1.881	0.013	1.898	0.611	0.617
BW −	All	240	75.5	65.1	31.0	63.1	5.790	5.077	2.785	0.908	0.449
BW −	Effective	239	75.4	65.0	30.9	63.2	5.805	5.090	2.792	0.910	0.449
BW −	Unchanged	1	NA	NA	75.5	NA	2.249	1.877	1.255	0.554	0.309
BW −	Opposite	—	—	—	—	—	—	—	—	—	—
RF +	All	240	45.3	39.6	91.4	85.6	0.829	-0.085	0.959	0.518	0.600
RF +	Effective	4	100.0	100.0	85.8	95.8	0.895	0.548	0.730	0.550	0.437
RF +	Unchanged	205	44.6	39.4	92.1	86.5	0.835	-0.054	0.950	0.526	0.601
RF +	Opposite	31	54.6	51.9	87.7	87.5	0.787	-0.373	1.049	0.463	0.615
RF −	All	240	42.6	36.6	90.9	85.0	0.849	-0.001	0.924	0.584	0.633
RF −	Effective	24	50.4	49.4	87.7	89.5	1.070	0.410	0.989	0.696	0.650
RF −	Unchanged	208	41.9	37.1	91.7	86.3	0.834	-0.026	0.919	0.579	0.634
RF −	Opposite	8	74.5	74.7	81.8	92.4	0.557	-0.584	0.857	0.365	0.558

Table 42: Outcome-group geometry and training-direction projections at Llama block 23. Template A; each row uses the same N_h activation-paired inputs for both sets of measurements. F_3/R_3 are full/residual top-three centered-variance percentages, each using its own branch denominator. E_T/E_C are retained residual percentages of original full total/centered energy. $\bar{p}$ is mean displacement projection in raw activation units onto a unit full, parallel, or residual training-fitted orientation; these separately normalized coordinates are not additive raw components. $\bar{c}$ averages cosines over nonzero displacements. All includes invalid-score pairs; response groups require valid paired scores. NA: undefined; —: empty or unavailable, not zero. BW/RF: Bandwagon/Refinement.

The final block preserves this distinction. Refinement− has mean parallel projections 1.287 and -0.086 in its effective and unchanged groups, while residual cosines remain close at 0.649 and 0.640. Absolute projections are expressed in raw activation units and change with depth; the within-block group comparisons and normalized cosines provide the relevant comparisons here. Favorable-cue effective groups contain only three or four test pairs, and their concentrated spectra should be read together with those displayed counts.

Cue	**Group**	N_h	F_3	R_3	E_T	E_C	$\bar{p}_F$	$\bar{p}_\parallel$	$\bar{p}_\perp$	$\bar{c}_F$	$\bar{c}_\perp$
BW +	All	240	55.1	49.3	81.3	83.9	4.580	-0.378	4.705	0.595	0.610
BW +	Effective	3	100.0	100.0	92.1	88.8	4.533	0.509	4.511	0.583	0.579
BW +	Unchanged	185	52.2	47.9	85.4	87.1	4.578	-0.212	4.676	0.615	0.627
BW +	Opposite	52	61.6	58.0	70.3	84.4	4.591	-1.017	4.822	0.526	0.551
BW −	All	240	76.3	67.7	32.1	65.9	16.545	14.631	7.727	0.885	0.422
BW −	Effective	239	76.2	67.7	32.0	66.1	16.594	14.675	7.747	0.886	0.423
BW −	Unchanged	1	NA	NA	82.6	NA	4.881	4.033	2.817	0.446	0.258
BW −	Opposite	—	—	—	—	—	—	—	—	—	—
RF +	All	240	43.2	37.4	91.1	84.7	2.686	-0.119	3.026	0.561	0.632
RF +	Effective	4	100.0	100.0	85.4	96.6	2.849	1.674	2.361	0.598	0.490
RF +	Unchanged	205	42.5	37.3	92.0	86.0	2.727	-0.015	3.023	0.572	0.637
RF +	Opposite	31	53.7	49.5	86.1	84.8	2.388	-1.039	3.131	0.480	0.622
RF −	All	240	41.2	34.4	90.2	84.2	2.598	-0.005	2.803	0.592	0.638
RF −	Effective	24	52.1	48.3	85.6	85.9	3.269	1.287	3.006	0.694	0.649
RF −	Unchanged	208	40.1	34.5	91.3	86.1	2.551	-0.086	2.784	0.588	0.640
RF −	Opposite	8	74.3	74.1	80.8	92.8	1.822	-1.784	2.682	0.385	0.565

Table 43: Outcome-group geometry and training-direction projections at Llama block 31. Template A; each row uses the same N_h activation-paired inputs for both sets of measurements. F_3/R_3 are full/residual top-three centered-variance percentages, each using its own branch denominator. E_T/E_C are retained residual percentages of original full total/centered energy. $\bar{p}$ is mean displacement projection in raw activation units onto a unit full, parallel, or residual training-fitted orientation; these separately normalized coordinates are not additive raw components. $\bar{c}$ averages cosines over nonzero displacements. All includes invalid-score pairs; response groups require valid paired scores. NA: undefined; —: empty or unavailable, not zero. BW/RF: Bandwagon/Refinement.

The effective-minus-unchanged contrasts quantify the differences along each fixed orientation. In both templates, Refinement's parallel projection contrasts are positive at all three observed blocks; the residual projection intervals include zero. At block 23, for example, the parallel contrasts are 0.436 for A and 0.457 for B, compared with residual contrasts of 0.070 and 0.018. Templates A and B form their groups from their own unsteered scores. These are comparisons between populations of different questions along the same training-fitted orientation.

Block	T	Direction	n_e	n_u	Δ projection	Δ cosine
15	A	Full	24	208	0.017 [-0.023, 0.054]	0.027 [0.003, 0.049]
15	A	Parallel (unit)	24	208	0.070 [0.038, 0.103]	0.077 [0.039, 0.114]
15	A	Residual (unit)	24	208	0.007 [-0.031, 0.043]	0.016 [-0.005, 0.037]
15	B	Full	20	208	0.012 [-0.026, 0.049]	0.009 [-0.012, 0.030]
15	B	Parallel (unit)	20	208	0.084 [0.046, 0.118]	0.083 [0.048, 0.115]
15	B	Residual (unit)	20	208	0.000 [-0.036, 0.036]	-0.003 [-0.025, 0.018]
23	A	Full	24	208	0.236 [0.141, 0.339]	0.117 [0.086, 0.148]
23	A	Parallel (unit)	24	208	0.436 [0.304, 0.581]	0.260 [0.182, 0.342]
23	A	Residual (unit)	24	208	0.070 [-0.002, 0.150]	0.016 [-0.014, 0.043]
23	B	Full	20	208	0.197 [0.099, 0.305]	0.107 [0.064, 0.150]
23	B	Parallel (unit)	20	208	0.457 [0.333, 0.577]	0.276 [0.213, 0.339]
23	B	Residual (unit)	20	208	0.018 [-0.064, 0.097]	-0.002 [-0.046, 0.036]
31	A	Full	24	208	0.718 [0.390, 1.036]	0.106 [0.070, 0.140]
31	A	Parallel (unit)	24	208	1.372 [0.913, 1.876]	0.260 [0.176, 0.349]
31	A	Residual (unit)	24	208	0.222 [-0.014, 0.454]	0.010 [-0.032, 0.044]
31	B	Full	20	208	0.633 [0.346, 0.922]	0.105 [0.071, 0.138]
31	B	Parallel (unit)	20	208	1.472 [1.064, 1.888]	0.280 [0.203, 0.356]
31	B	Residual (unit)	20	208	0.091 [-0.176, 0.340]	0.001 [-0.039, 0.037]

Table 44: Refinement-negative effective-minus-unchanged contrasts along training-fitted directions. Projection uses raw activation units; cosine normalizes each displacement. Full and unit components use the expanded Llama clean-score reference of rank at most eleven. Groups contain different questions, not paired effective/unchanged counterfactuals. T denotes template; n_e/n_u are effective/unchanged group sizes. Brackets are pointwise 95% benchmark-stratified question-cluster intervals.

The matched-arm analysis asks a separate question: how does the broader reference decompose changes caused by framing and final-answer errors at identical score-change strata? At block 23, the Bandwagon framing and content arms retain 30.8% and 39.8% of their original total displacement energy outside the reference. The residuals still have concentrated centered variation, with their own first three PCs explaining 66.0% and 61.3%. Thus, removing the measured score reference changes both the size and the spectrum of the displacement without eliminating its low-dimensional concentration. The full-direction mean alignments and rank-two comparison appear in Appendix G.2; the small Refinement matched set is displayed with its sample size.

Cue/arm	Block	N_h	Full top-3 (%)	Residual top-3 (%)	Retained total (%)	Retained centered (%)
BW −/F	15	175	50.0	46.9	62.8	88.8
BW −/C	15	175	50.4	48.5	70.1	92.2
BW −/F	23	175	76.3	66.0	30.8	63.1
BW −/C	23	175	74.1	61.3	39.8	61.4
BW −/F	31	175	77.3	69.0	31.8	66.0
BW −/C	31	175	73.7	60.2	39.2	60.0
RF −/F	15	5	86.8	86.6	97.6	97.8
RF −/C	15	5	85.9	86.2	72.2	92.9
RF −/F	23	5	91.1	90.9	85.6	95.6
RF −/C	23	5	93.0	91.7	45.0	79.8
RF −/F	31	5	90.3	89.7	83.1	90.4
RF −/C	31	5	94.1	93.0	49.8	79.4

Table 45: Exact-matched test displacement geometry across the three observed blocks. F: framing; C: final-answer corruption. Both arms have identical benchmark, baseline-score, and score-drop distributions under exact matching. Each spectrum divides by its own centered energy; retained fractions divide residual energy by the original full energy. RF is descriptive under the sample-support criteria. BW: bandwagon; RF: refinement. NA denotes an undefined statistic; — denotes an unavailable method or empty group, not zero.

J Interventions with the Logit-Augmented Reference

The Llama interventions test whether the geometry under the broader reference corresponds to changes in emitted ratings. Forward interventions cover favorable and unfavorable framing directions; reverse interventions test movement from unfavorable framed ratings toward each answer's clean reference. The parallel and residual directions use the logit-augmented span from Appendix E.1. The common-reference forward comparison appears in Appendix H.1.

J.1 Primary and Strict-Format Endpoints

Both scoring policies evaluate the same fitted directions, development-selected strengths, and generated completions. Response-group membership follows the primary paired scores. Policy-specific validity counts and valid-pair estimates show which ratings satisfy the stricter whole-response rule.

The conservative endpoints average over all inputs under the missing-reference convention in Appendix E.5, rather than the fixed valid references used in Appendices H.1 and K. Consequently, a negative gain at zero strength can reflect a missing original reference rather than an intervention effect; valid-triple gain and valid-pair collateral are reported with their coverage.

J.2 Forward Configurations and Test Effects

The comparison includes all four framing conditions, their score-reference components, clean-score and neutral directions, five random draws, and available matched-source directions. Each unit orientation receives the same signed development grid. Original-magnitude components instead keep the full direction's coefficient, so their development feasibility belongs to that full-direction selection. Their measured test effects and rank correlations describe the component interventions themselves.

The tables place development settings beside the corresponding test effects, so the strength of each intervention can be read together with the ranking and validity it preserves. Signed shift is the steered rating minus the clean rating. Polarity-aligned shift follows the condition's intended increase or decrease, with invalid endpoints receiving a negative penalty in its conservative mean. The favorable directions use eight effective Bandwagon training pairs and six Refinement pairs; these entries provide the low-support polarity comparison under the stated sample-count rule.

For Bandwagon−, the full and unit parallel directions each produce a 0.117-point primary decrease, while the original-magnitude residual produces only 0.004. The independently normalized residual reaches a 0.058-point decrease at a larger selected coefficient and maintains rank correlation 0.948. Reading the configuration and effect together distinguishes the small original residual component from the score response available along its rescaled orientation.

±	Direction	r/α	V_d/ρ_d	N/N_3	V/ρ	Δs	Aligned shift [95% CI]
+	Full	0.160 2.802	100.0 0.913 ✓	240/240 240/161	100.0/0.944 67.1/0.864	0.033 0.075	0.033 [-0.004, 0.067] -2.913 [-3.108, -2.712]
+	Parallel	-0.080 -1.401	100.0 0.865 ✓	240/240 240/159	100.0/0.928 66.2/0.815	0.008 0.088	0.008 [-0.042, 0.054] -2.979 [-3.175, -2.792]
+	Residual	0.160 2.802	100.0 0.869 ✓	240/240 240/160	100.0/0.940 66.7/0.853	0.037 0.081	0.037 [-0.004, 0.075] -2.946 [-3.154, -2.733]
+	Neutral	-0.040 -0.700	100.0 0.880 ✓	240/240 240/162	100.0/0.873 67.5/0.838	0.158 0.099	0.158 [0.104, 0.221] -2.858 [-3.071, -2.638]
+	Parallel (raw)	0.160 2.802	100.0 0.913 ✓	240/240 240/162	100.0/0.990 67.5/0.967	0.000 0.000	0.000 [-0.017, 0.021] -2.925 [-3.121, -2.717]
+	Residual (raw)	0.160 2.802	100.0 0.913 ✓	240/240 240/160	100.0/0.940 66.7/0.853	0.037 0.081	0.037 [-0.004, 0.075] -2.946 [-3.154, -2.733]
−	Full	0.020 0.350	100.0 0.862 ✓	240/240 240/160	100.0/0.919 67.1/0.934	-0.117 -0.025	0.117 [0.067, 0.171] -2.983 [-3.183, -2.787]
−	Parallel	0.020 0.350	100.0 0.862 ✓	240/240 240/159	100.0/0.919 66.7/0.932	-0.117 -0.025	0.117 [0.067, 0.171] -3.021 [-3.213, -2.837]
−	Residual	0.040 0.700	100.0 0.894 ✓	240/240 240/161	100.0/0.948 67.1/0.935	-0.058 -0.025	0.058 [0.025, 0.096] -2.946 [-3.133, -2.754]
−	Neutral	0.040 0.700	100.0 0.851 ✓	240/240 240/159	100.0/0.919 66.7/0.932	-0.117 -0.025	0.117 [0.067, 0.171] -3.021 [-3.213, -2.837]
−	Parallel (raw)	0.020 0.350	100.0 0.862 ✓	240/240 240/159	100.0/0.924 66.2/0.932	-0.100 -0.025	0.100 [0.050, 0.154] -3.021 [-3.213, -2.837]
−	Residual (raw)	0.020 0.350	100.0 0.862 ✓	240/240 240/162	100.0/0.995 67.5/0.984	-0.004 -0.006	0.004 [0.000, 0.013] -2.921 [-3.112, -2.729]
−	Matched frame	0.020 0.350	100.0 0.862 ✓	240/240 240/160	100.0/0.919 67.1/0.934	-0.117 -0.025	0.117 [0.067, 0.171] -2.983 [-3.183, -2.787]
−	Matched content	0.020 0.350	100.0 0.880 ✓	240/240 240/160	100.0/0.935 66.7/0.951	-0.096 -0.019	0.096 [0.050, 0.146] -2.987 [-3.183, -2.792]

Table 46: Bandwagon-derived and cue-matched forward controls with the expanded Llama reference. For test columns, the upper/lower line is first-line/strict whole-response scoring of the same completion. Development uses primary scores: r/α lists relative/absolute strength; V_d/ρ_d lists validity (%) and rank correlation, with ✓/× for feasible/infeasible selection ($V_d \geq 93\%$, $\rho_d \geq 0.85$). Raw components inherit the full coefficient and development metrics; unit components are independently searched. N/N_3: all inputs/valid clean–original–steered triples. Test V/ρ: validity (%)/valid-pair rank correlation. Signed shift is steered minus clean; aligned shift follows polarity and assigns -9 to invalid endpoints. Brackets are 95% benchmark-stratified question-cluster intervals conditional on training and development selection. Favorable directions use low effective-training support. Matched frame/content are exact-matched training-arm means; the neutral direction uses the corresponding cue family.

Refinement supplies the complementary comparison because most of its mean-direction energy lies outside the expanded reference. Its component interventions therefore test whether geometric size and score effect track one another under a different allocation of energy. The table includes the cue-matched neutral direction and both matched-source directions alongside the full, raw-component, and unit-component interventions. Favorable and unfavorable settings remain separate because their training populations and intended score shifts differ.

±	Direction	r/α	V_d/ρ_d	N/N_3	V/ρ	Δs	**Aligned shift [95% CI]**
+	Full	0.080 1.401	100.0 0.905 ✓	240/240 240/162	100.0/0.915 67.5/0.958	0.104 0.031	0.104 [0.054, 0.158] -2.904 [-3.112, -2.692]
+	Parallel	0.020 0.350	100.0 0.889 ✓	240/240 240/162	100.0/0.885 67.5/0.838	0.142 0.099	0.142 [0.092, 0.200] -2.858 [-3.071, -2.638]
+	Residual	-0.080 -1.401	100.0 0.878 ✓	240/240 240/160	100.0/0.903 66.7/0.810	0.117 0.113	0.117 [0.071, 0.171] -2.925 [-3.138, -2.708]
+	Neutral	-0.040 -0.700	100.0 0.855 ✓	240/240 240/162	100.0/0.863 67.5/0.827	0.171 0.105	0.171 [0.117, 0.233] -2.854 [-3.067, -2.638]
+	Parallel (raw)	0.080 1.401	100.0 0.905 ✓	240/240 240/162	100.0/0.845 67.5/0.814	0.196 0.117	0.196 [0.138, 0.263] -2.846 [-3.058, -2.633]
+	Residual (raw)	0.080 1.401	100.0 0.905 ✓	240/240 240/161	100.0/0.907 67.1/0.784	-0.096 -0.081	-0.096 [-0.142, -0.054] -3.017 [-3.204, -2.825]
−	Full	0.040 0.700	100.0 0.851 ✓	240/240 240/161	100.0/0.886 67.5/0.818	-0.146 -0.068	0.146 [0.087, 0.204] -2.917 [-3.104, -2.725]
−	Parallel	0.020 0.350	100.0 0.862 ✓	240/240 240/160	100.0/0.914 67.1/0.918	-0.121 -0.031	0.121 [0.067, 0.179] -2.979 [-3.179, -2.779]
−	Residual	0.080 1.401	100.0 0.911 ✓	240/240 240/161	100.0/0.928 67.1/0.818	-0.071 -0.068	0.071 [0.037, 0.113] -2.917 [-3.104, -2.725]
−	Neutral	0.020 0.350	100.0 0.899 ✓	240/240 240/160	100.0/0.958 66.7/0.967	-0.058 -0.013	0.058 [0.021, 0.100] -2.992 [-3.188, -2.796]
−	Parallel (raw)	0.040 0.700	100.0 0.851 ✓	240/240 240/160	100.0/0.924 66.7/0.934	-0.100 -0.025	0.100 [0.050, 0.154] -2.983 [-3.183, -2.787]
−	Residual (raw)	0.040 0.700	100.0 0.851 ✓	240/240 240/162	100.0/0.981 67.5/0.935	-0.017 -0.025	0.017 [0.004, 0.033] -2.908 [-3.104, -2.712]
−	Matched frame	0.040 0.700	100.0 0.857 ✓	240/240 240/161	100.0/0.895 67.5/0.852	-0.138 -0.056	0.138 [0.083, 0.196] -2.925 [-3.112, -2.733]
−	Matched content	0.020 0.350	100.0 0.875 ✓	240/240 240/160	100.0/0.924 66.7/0.934	-0.100 -0.025	0.100 [0.050, 0.154] -2.983 [-3.183, -2.787]

Table 47: Refinement-derived and cue-matched forward controls with the expanded Llama reference. For test columns, the upper/lower line is first-line/strict whole-response scoring of the same completion. Development uses primary scores: r/α lists relative/absolute strength; V_d/ρ_d lists validity (%) and rank correlation, with ✓/× for feasible/infeasible selection ($V_d \geq 93\%$, $\rho_d \geq 0.85$). Raw components inherit the full coefficient and development metrics; unit components are independently searched. N/N_3: all inputs/valid clean–original–steered triples. Test V/ρ: validity (%)/valid-pair rank correlation. Signed shift is steered minus clean; aligned shift follows polarity and assigns -9 to invalid endpoints. Brackets are 95% benchmark-stratified question-cluster intervals conditional on training and development selection. Favorable directions use low effective-training support. Matched frame/content are exact-matched training-arm means; the neutral direction uses the corresponding cue family.

The clean-score and random references have identical selected configurations and test effects for the two cue families within each polarity, so Table 48 displays each shared comparison once. Their strengths are selected on the same development grid as the unit framing directions. Under the unfavorable objective, the clean high–low direction yields the same 0.117-point decrease as the Bandwagon full direction; the five random draws span smaller primary decreases from 0.008 to 0.058 points. These comparisons place the cue-derived effects on the same score and rank-correlation scale as directions obtained without the cue.

±	Direction	r/α	V_d/ρ_d	N/N_3	V/ρ	Δs	Aligned shift [95% CI]
+	High-low	0.020 0.350	100.0 0.889 ✓	240/240 240/162	100.0/0.879 67.5/0.838	0.150 0.099	0.150 [0.096, 0.208] -2.858 [-3.071, -2.638]
+	Ridge	0.080 1.401	100.0 0.874 ✓	240/240 240/161	100.0/0.879 67.1/0.838	0.150 0.099	0.150 [0.096, 0.208] -2.896 [-3.117, -2.671]
+	Random 1	-0.010 -0.175	100.0 0.989 ✓	240/240 240/162	100.0/1.000 67.5/1.000	0.000 0.000	0.000 [0.000, 0.000] -2.925 [-3.112, -2.737]
+	Random 2	-0.160 -2.802	100.0 0.883 ✓	240/240 240/160	100.0/0.929 66.7/0.853	0.013 0.069	0.013 [-0.029, 0.054] -2.954 [-3.158, -2.746]
+	Random 3	0.160 2.802	100.0 0.913 ✓	240/240 240/160	100.0/0.952 66.7/0.848	0.054 0.081	0.054 [0.025, 0.087] -2.946 [-3.150, -2.733]
+	Random 4	0.000 0.000	100.0 1.000 ✓	240/240 240/162	100.0/1.000 67.5/1.000	0.000 0.000	0.000 [0.000, 0.000] -2.925 [-3.112, -2.737]
+	Random 5	0.160 2.802	100.0 0.914 ✓	240/240 240/162	100.0/0.928 67.5/0.958	0.087 0.031	0.087 [0.042, 0.142] -2.904 [-3.112, -2.692]
−	High-low	-0.020 -0.350	100.0 0.862 ✓	240/240 240/159	100.0/0.919 66.7/0.932	-0.117 -0.025	0.117 [0.067, 0.171] -3.021 [-3.213, -2.837]
−	Ridge	-0.080 -1.401	100.0 0.850 ✓	240/240 240/161	100.0/0.919 67.5/0.935	-0.100 -0.025	0.100 [0.050, 0.154] -2.946 [-3.133, -2.754]
−	Random 1	0.160 2.802	100.0 0.938 ✓	240/240 240/160	100.0/0.968 66.7/0.915	-0.013 -0.025	0.013 [-0.013, 0.037] -2.983 [-3.175, -2.796]
−	Random 2	0.160 2.802	100.0 0.901 ✓	240/240 240/162	100.0/0.863 67.5/0.669	-0.008 -0.074	0.008 [-0.046, 0.058] -2.875 [-3.071, -2.679]
−	Random 3	-0.160 -2.802	100.0 0.916 ✓	240/240 240/162	100.0/0.957 67.5/0.852	-0.046 -0.056	0.046 [0.021, 0.075] -2.888 [-3.083, -2.683]
−	Random 4	0.160 2.802	100.0 0.941 ✓	240/240 240/161	100.0/0.949 67.1/0.954	-0.042 0.019	0.042 [0.000, 0.083] -2.975 [-3.158, -2.788]
−	Random 5	-0.160 -2.802	100.0 0.916 ✓	240/240 240/160	100.0/0.958 66.7/0.966	-0.058 -0.013	0.058 [0.021, 0.100] -2.992 [-3.175, -2.829]

Table 48: Shared clean-score and random forward controls with the expanded Llama reference. For test columns, the upper/lower line is first-line/strict whole-response scoring of the same completion. Development uses primary scores: r/α lists relative/absolute strength; V_d/ρ_d lists validity (%) and rank correlation, with ✓/× for feasible/infeasible selection ($V_d \geq 93\%$, $\rho_d \geq 0.85$). Raw components inherit the full coefficient and development metrics; unit components are independently searched. N/N_3: all inputs/valid clean–original–steered triples. Test V/ρ: validity (%)/valid-pair rank correlation. Signed shift is steered minus clean; aligned shift follows polarity and assigns -9 to invalid endpoints. Brackets are 95% benchmark-stratified question-cluster intervals conditional on training and development selection. The shared settings and endpoints are identical for Bandwagon and Refinement within each polarity. Random 1–5 use seeds 101–105.

J.3 Reverse Selection, Paired Error, and Collateral Effects

No nonrandom direction met the reverse-development constraints on the signed grid. Those methods therefore use zero-strength fallbacks. Random 2, corresponding to seed 102, is the nonzero Refinement− exception. Its test paired-MAE gain is 0.025 points on template A (95% CI $[0.000, 0.058]$) and 0.000 on B ($[-0.029, 0.033]$). Test rank correlations are 0.849 and 0.826, both below the 0.85 fidelity threshold.

Before intervention, development correlations with clean scores are 0.280 for Bandwagon− and 0.806 for Refinement−, already below that threshold. Feasibility thus requires improved ranking as well as reduced paired error. A uniform upward score shift can reduce the latter but cannot improve

rank correlation. The response curves in Appendix K examine these outcomes across the full sampled strength range.

Table 49 connects these selection outcomes to paired recovery on both framed templates. Paired gain compares absolute errors to each question's clean reference. Closer and worse rates use valid triples; exact recovery and overshoot use the subset whose original framed score differs from clean. For the nonzero Refinement configuration, 2.1% of primary template-A scores move closer and 0.4% move farther, whereas template B has 0.8% closer and 1.2% farther. The distinction between a small mean gain and the fraction of affected questions is therefore visible in the same comparison.

Development-selected configurations

Cue	Directions	r/α	V_d/ρ_d
BW −	All directions	0.000 0.000	100.0 0.280 ×
RF −	All except Random 2	0.000 0.000	100.0 0.806 ×
RF −	Random 2	-0.080 -1.401	100.0 0.850 ✓

Framed templates A and B

Input/config.	$N/N_3/N_d$	V/ρ	MAE / gain$_v$	Gain [95% CI]	Closer/worse/ exact/overshoot (%)
BW − A	240/240/239	100.0/0.376	3.592/0.000	0.000 [0.000, 0.000]	0.0/0.0/0.0/0.0
Zero	240/114/113	54.6/0.430	3.842/0.000	-4.725 [-5.175, -4.312]	0.0/0.0/0.0/0.0
BW − B	240/240/239	100.0/0.461	2.996/0.000	0.000 [0.000, 0.000]	0.0/0.0/0.0/0.0
Zero	240/153/152	78.8/0.339	3.176/0.000	-3.263 [-3.525, -3.038]	0.0/0.0/0.0/0.0
RF − A	240/240/32	100.0/0.821	0.204/0.000	0.000 [0.000, 0.000]	0.0/0.0/0.0/0.0
Zero	240/160/16	69.2/0.731	0.106/0.000	-3.000 [-3.188, -2.812]	0.0/0.0/0.0/0.0
RF − A	240/240/32	100.0/0.849	0.204/0.025	0.025 [0.000, 0.058]	2.1/0.4/15.6/0.0
Random 2	240/160/16	69.2/0.771	0.106/0.013	-2.992 [-3.188, -2.788]	1.9/0.6/18.8/0.0
RF − B	240/240/32	100.0/0.824	0.208/0.000	0.000 [0.000, 0.000]	0.0/0.0/0.0/0.0
Zero	240/160/15	69.2/0.755	0.100/0.000	-3.000 [-3.188, -2.850]	0.0/0.0/0.0/0.0
RF − B	240/240/32	100.0/0.826	0.208/0.000	0.000 [-0.029, 0.033]	0.8/1.2/6.2/0.0
Random 2	240/160/15	68.8/0.724	0.100/-0.013	-3.008 [-3.192, -2.850]	0.0/1.2/0.0/0.0

Table 49: Reverse-selection feasibility and paired recovery in Llama. Upper/lower test lines are primary/strict scores at the same selected setting. The direction set is full, unit parallel/residual, clean high–low, ridge, neutral, Random 1–5 (seeds 101–105), raw parallel/residual, and matched frame/content. V_d/ρ_d is development validity (%) above correlation; marks use the $93\%/0.85$ criteria. All nonexceptional settings are zero-strength fallbacks. N_3 counts valid triples; N_d counts originally changed triples. MAE is original paired absolute error; gain$_v$ uses valid triples. Conservative gain penalizes invalid steered scores with loss 9 and missing original references with gain -9. Thus strict negative gain at zero includes missing-reference penalties. Closer/worse use N_3; exact/overshoot use N_d. V/ρ gives test validity (%)/valid-pair rank correlation. Intervals use 2,000 benchmark-stratified question-cluster draws, conditional on the fixed training fit and primary development selection. BW/RF: Bandwagon/Refinement.

The same reverse settings are also applied to clean and content-altered answers. The nonzero Refinement configuration changes their primary ratings by mean absolute amounts of 0.058 and 0.023 points. Their jointly steered clean-minus-content gap is 3.579 points, compared with 3.565 before intervention. The fixed-clean gap instead changes only the content score, separating that response from a simultaneous change to the clean rating. Table 50 keeps the valid-pair and valid-triple populations distinct and reports the corresponding strict-format endpoints. Zero-reference rows are shared across cues only where the same no-intervention scores yield identical estimates.

Paired error and absolute score change

Input/ config.	$N/N_3/N_v$	V/ρ	Gain$_v$	Gain [95% CI]	Absolute change [CI] / valid-pair mean
Clean	240/240/240	100.0/1.000	0.000	0.000 [0.000, 0.000]	0.000 [0.000, 0.000] / 0.000
Zero	240/162/162	67.5/1.000	0.000	-2.925 [-3.112, -2.737]	2.925 [2.737, 3.112] / 0.000
Content	214/214/214	100.0/-0.059	0.000	0.000 [0.000, 0.000]	0.000 [0.000, 0.000] / 0.000
Zero	214/91/107	50.0/0.149	0.000	-5.173 [-5.636, -4.752]	4.500 [3.995, 5.005] / 0.000
Clean	240/240/240	100.0/0.944	-0.058	-0.058 [-0.096, -0.029]	0.058 [0.029, 0.096] / 0.058
RF R2	240/160/160	66.7/0.864	-0.062	-3.042 [-3.246, -2.842]	3.042 [2.842, 3.246] / 0.062
Content	214/214/214	100.0/-0.057	0.014	0.014 [-0.009, 0.047]	0.023 [0.000, 0.056] / 0.023
RF R2	214/90/106	49.5/0.144	0.033	-5.178 [-5.626, -4.748]	4.565 [4.042, 5.089] / 0.047

Same-question clean-minus-content gaps

Config.	$N_o/N_f/N_j$	Original gap [CI]	Fixed-clean gap [CI]	Joint gap [CI]
Zero	214/214/214	3.565 [3.383, 3.738]	3.565 [3.383, 3.738]	3.565 [3.383, 3.738]
(both cues)	91/91/91	4.242 [3.968, 4.536]	4.242 [3.968, 4.536]	4.242 [3.968, 4.536]
RF	214/214/214	3.565 [3.383, 3.738]	3.551 [3.374, 3.720]	3.579 [3.406, 3.738]
Random 2	91/90/90	4.242 [3.968, 4.536]	4.200 [3.935, 4.483]	4.233 [3.962, 4.521]

Table 50: Clean and content-altered ratings under the selected reverse settings. Upper/lower lines are primary/strict endpoints. The common zero row covers both cues; R2 is Refinement Random 2 (seed 102). N_3 counts valid clean/original/steered triples, while N_v counts valid original/steered pairs; these denominators remain distinct. V/ρ is output validity (%)/valid clean–steered rank correlation. Gain$_v$ is valid-triple paired MAE gain. Conservative gain uses the missing-reference convention in Appendix E.5; conservative absolute change assigns 9 to an invalid endpoint. The final column gives that conservative estimate and CI, followed by the valid-pair mean. Strict zero-reference penalties are not changes caused by intervention. Original gaps use both unsteered scores; fixed-clean gaps steer content only; joint gaps steer both scores, with separate valid-only counts N_o, N_f, N_j. Intervals use 2,000 benchmark-stratified question-cluster draws conditional on the fixed fit and primary development selection.

K LLAMA RESPONSE CURVES

K.1 STRENGTH DEPENDENCE BEYOND A SINGLE CONFIGURATION

The reverse-selection result combines two requirements: reducing paired error and preserving ranking. To examine how these outcomes vary with intervention strength, we evaluate the complete 13-strength signed grid for the same Llama directions and test populations. This descriptive analysis was specified after inspecting the selected-configuration results. It measures score responses across the grid without selecting a new configuration from test outcomes.

The four directions are each negative cue's full mean, its unit parallel and residual components, and the clean high–low reference. Components use the logit-augmented span of rank at most eleven. Each direction is applied to clean answers, framed templates A and B, and content-altered answers. Clean and each framed population contain 240 test questions; the content-altered population contains the 214 answers whose final field initially matches the benchmark reference. Curves use relative strength r from Eq. (7). The high–low direction points toward higher clean scores; framing-derived directions keep the signs of their effective training means.

Each scoring policy fixes its initially valid references before comparing strengths, using the endpoints in Appendix E.5. Strict-format curves evaluate the same directions, strengths, and completions. Bands are pointwise 95% benchmark-stratified question-cluster intervals, and connecting segments join the measured settings. The $\rho_S = 0.85$ line marks the fidelity requirement used for development selection; it is a reference for these test curves rather than a new selection rule.

K.2 RECOVERY AND COLLATERAL RESPONSE CURVES

Paired gain measures whether framed ratings move toward their clean references, while rank correlation measures preservation of question ordering. The response panels show both quantities separately for templates A and B. The collateral panels measure absolute changes to clean and content-altered ratings, changes in their jointly steered score gap, and the fractions moving toward the clean baseline. Together, these outcomes distinguish recovery of clean-reference scores from changes to the judge's other ratings across the full sampled strength range.

For Bandwagon, the negative full-direction endpoint moves most framed ratings closer to their clean references: template-A paired gain is 1.996 points and the toward-clean fraction is 95.8%. The response panels place that improvement beside its rank correlation of 0.203, while the collateral panels show the simultaneous changes to clean and content-altered ratings. In particular, the jointly steered clean-minus-content gap falls from 3.565 to 1.944 points. Following the same strength across the two figures therefore relates apparent recovery on the framed inputs to changes in the judge's other evaluations.

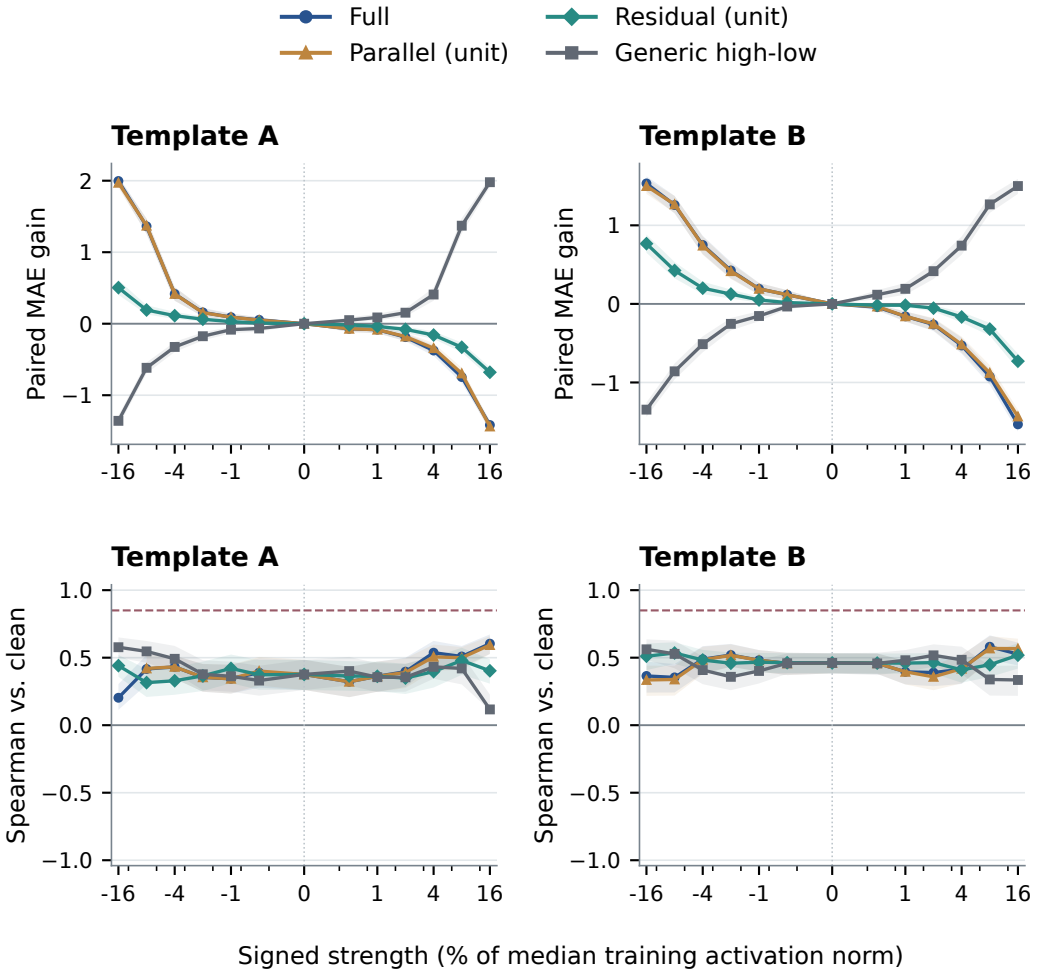


Figure 6: Bandwagon-negative response curves under first-line scoring. Top: paired MAE gain toward each answer's clean score for templates A and B; positive is improvement. Bottom: clean-versus-steered Spearman; the dashed line marks the 0.85 ranking reference. All 13 signed strengths are shown on a symmetric-log x-axis (linear within 0.5%); colors/markers distinguish full, unit parallel, unit residual, and generic high–low directions in Llama block 23 using the expanded 11-dimensional reference. Shading is pointwise 95% benchmark-stratified question-cluster uncertainty, not simultaneous bands. No test strength is selected. Fixed supports: clean 240, A 240, B 240, content 214. Later invalid scores incur loss 9 for MAE/collateral/movement; rank and signed gaps use valid endpoints. Undefined rank values are not imputed or connected across missing points.

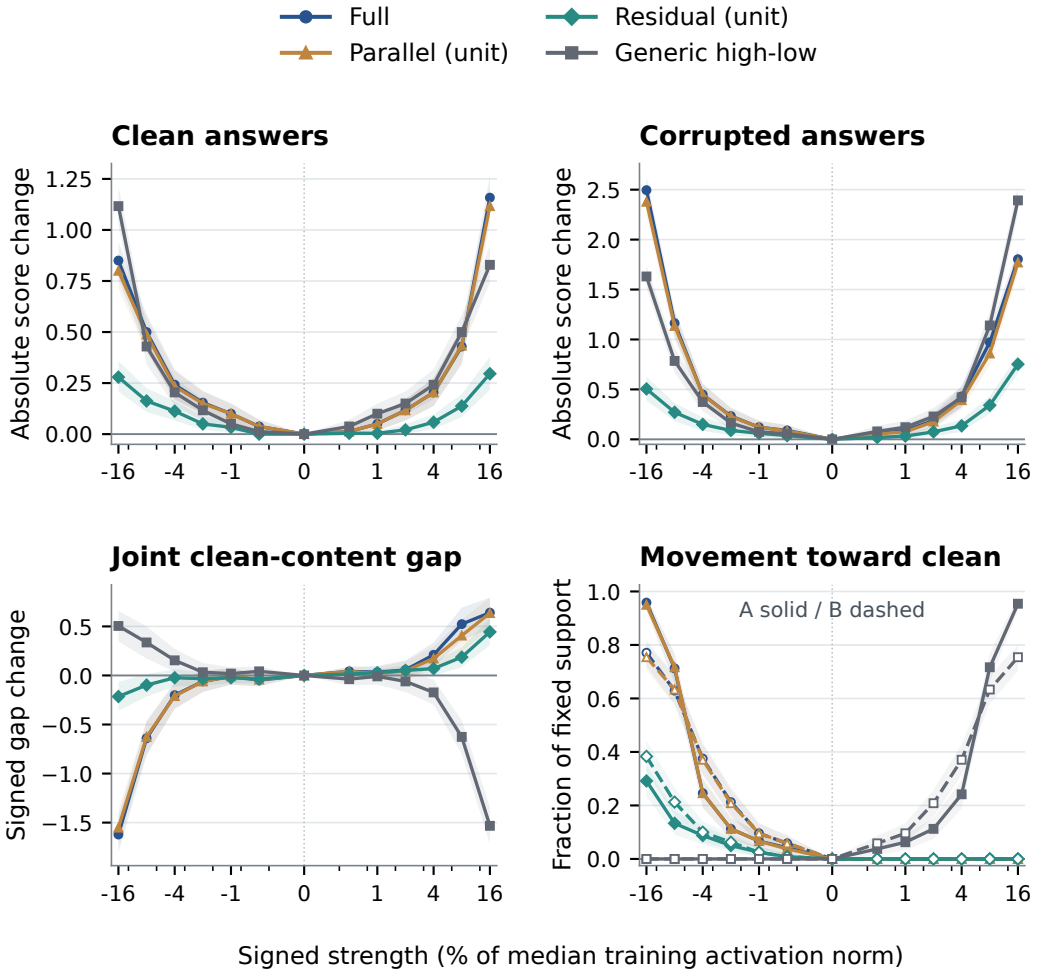


Figure 7: Bandwagon-negative collateral and movement curves under first-line scoring. Absolute clean/content score changes quantify collateral. The signed joint-gap change compares jointly steered clean-minus-content scores with the original same-question gap. The final panel gives toward-clean fractions (A solid, B dashed) on fixed supports. All 13 signed strengths are shown on a symmetric-log x-axis (linear within 0.5%); colors/markers distinguish full, unit parallel, unit residual, and generic high–low directions in Llama block 23 using the expanded 11-dimensional reference. Shading is pointwise 95% benchmark-stratified question-cluster uncertainty, not simultaneous bands. No test strength is selected. Fixed supports: clean 240, A 240, B 240, content 214. Later invalid scores incur loss 9 for MAE/collateral/movement; rank and signed gaps use valid endpoints. Undefined rank values are not imputed or connected across missing points.

Refinement begins from much smaller framing-induced errors, making the component comparison especially informative. At the negative endpoint, its full direction worsens paired MAE by 0.313 points, whereas its unit residual has estimated gain 0.033 with interval $[-0.021, 0.088]$. The latter retains rank correlation 0.865 but also changes clean and content-altered ratings. The four panels distinguish these quantities throughout the sampled grid: a small paired improvement, preservation of ordering, and small collateral changes are separate outcomes rather than interchangeable measures of recovery.

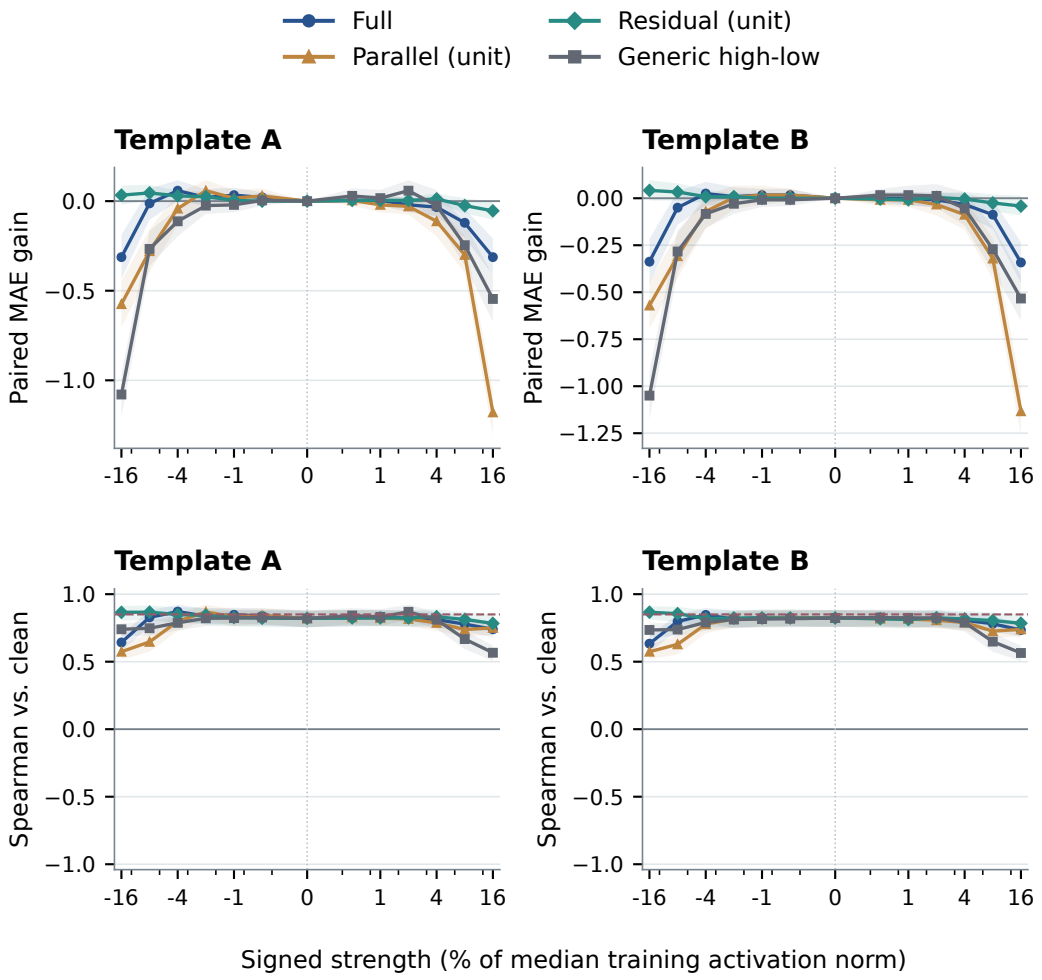


Figure 8: Refinement-negative response curves under first-line scoring. Top: paired MAE gain toward each answer's clean score for templates A and B; positive is improvement. Bottom: clean-versus-steered Spearman; the dashed line marks the 0.85 ranking reference. All 13 signed strengths are shown on a symmetric-log x-axis (linear within 0.5%); colors/markers distinguish full, unit parallel, unit residual, and generic high–low directions in Llama block 23 using the expanded 11-dimensional reference. Shading is pointwise 95% benchmark-stratified question-cluster uncertainty, not simultaneous bands. No test strength is selected. Fixed supports: clean 240, A 240, B 240, content 214. Later invalid scores incur loss 9 for MAE/collateral/movement; rank and signed gaps use valid endpoints. Undefined rank values are not imputed or connected across missing points.

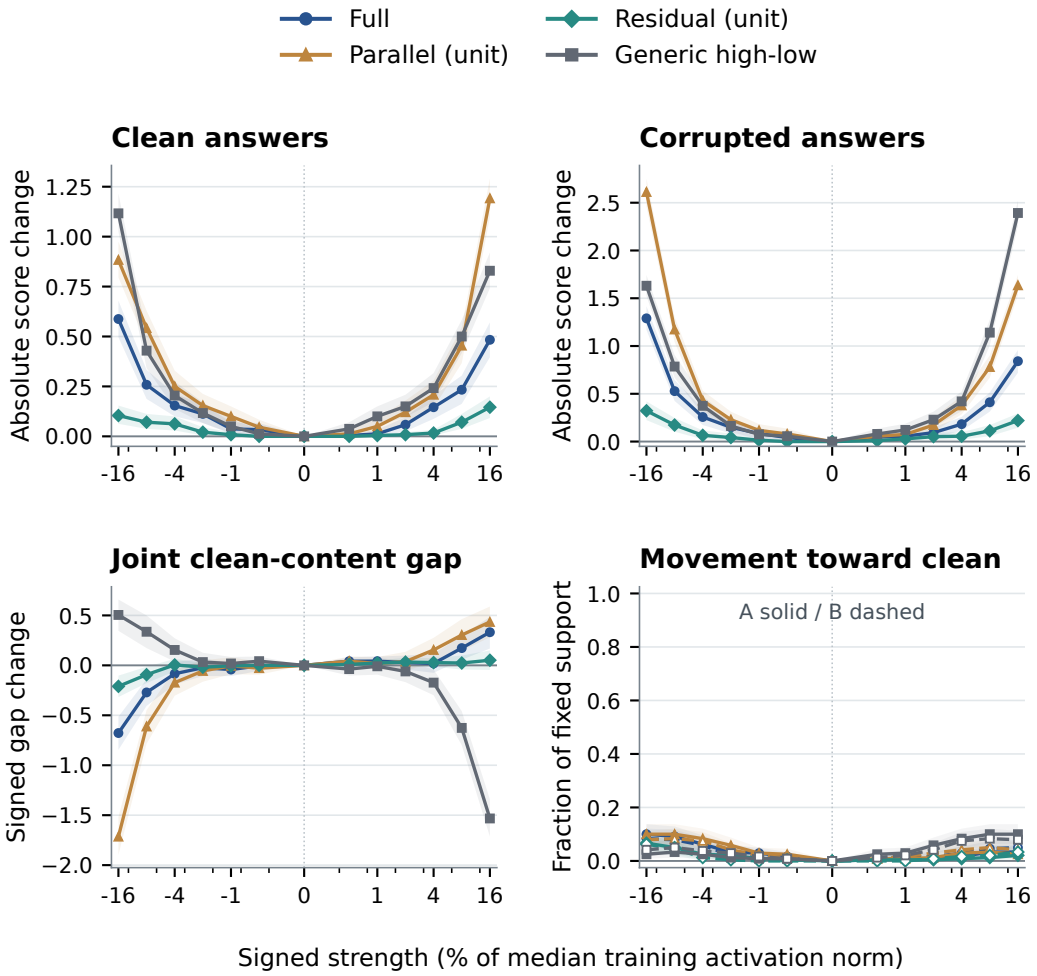


Figure 9: Refinement-negative collateral and movement curves under first-line scoring. Absolute clean/content score changes quantify collateral. The signed joint-gap change compares jointly steered clean-minus-content scores with the original same-question gap. The final panel gives toward-clean fractions (A solid, B dashed) on fixed supports. All 13 signed strengths are shown on a symmetric-log x-axis (linear within 0.5%); colors/markers distinguish full, unit parallel, unit residual, and generic high–low directions in Llama block 23 using the expanded 11-dimensional reference. Shading is pointwise 95% benchmark-stratified question-cluster uncertainty, not simultaneous bands. No test strength is selected. Fixed supports: clean 240, A 240, B 240, content 214. Later invalid scores incur loss 9 for MAE/collateral/movement; rank and signed gaps use valid endpoints. Undefined rank values are not imputed or connected across missing points.

The strict-format curves apply the whole-response rule to the same completions. For Bandwagon, the initially valid framed supports contain 114 template-A and 153 template-B questions, compared with 240 each under first-line scoring. Fixing these supports before varying strength separates the score response on compliant references from the selection of a different valid population at every grid point. Subsequent invalid outputs still contribute the specified loss to error, collateral, and movement measurements; correlation and signed gaps use their valid endpoints.

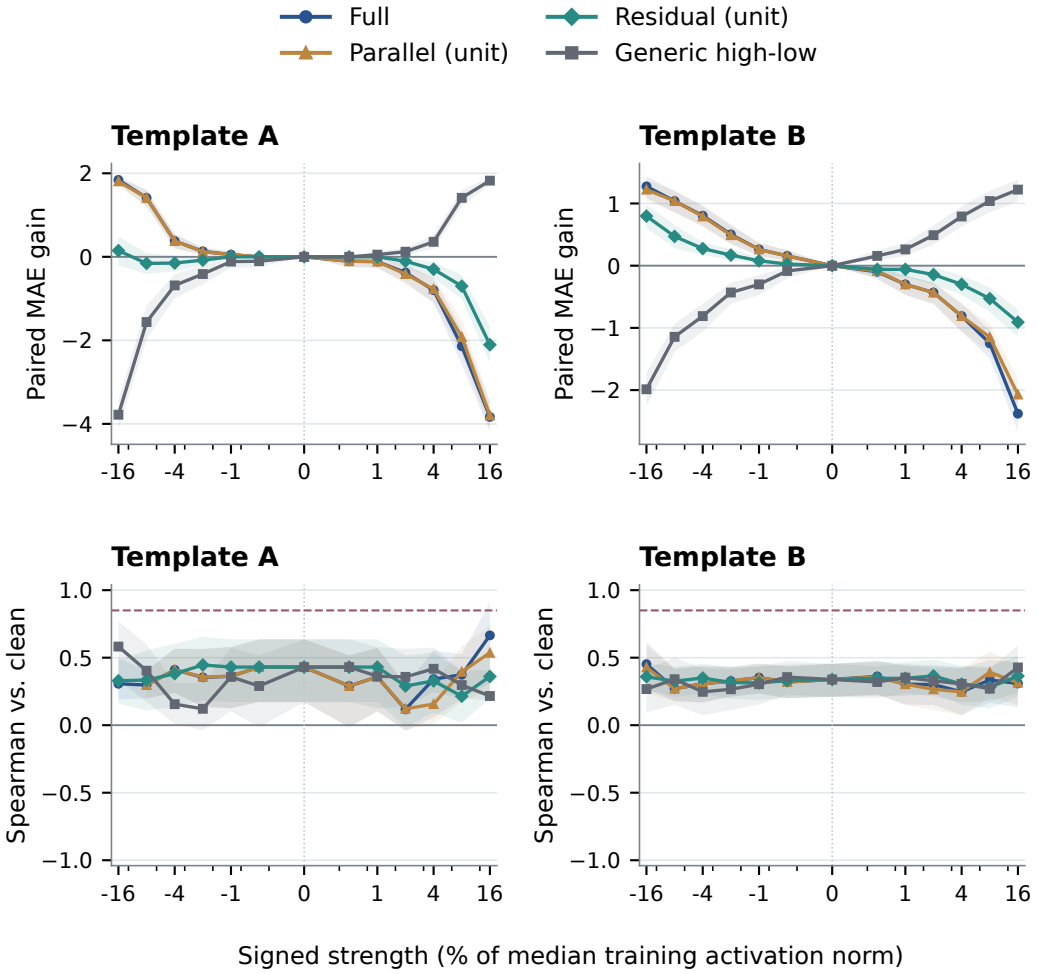


Figure 10: Bandwagon-negative response curves under strict whole-response scoring. Top: paired MAE gain toward each answer's clean score for templates A and B; positive is improvement. Bottom: clean-versus-steered Spearman; the dashed line marks the 0.85 ranking reference. All 13 signed strengths are shown on a symmetric-log x-axis (linear within 0.5%); colors/markers distinguish full, unit parallel, unit residual, and generic high–low directions in Llama block 23 using the expanded 11-dimensional reference. Shading is pointwise 95% benchmark-stratified question-cluster uncertainty, not simultaneous bands. No test strength is selected. Fixed supports: clean 162, A 114, B 153, content 91. Later invalid scores incur loss 9 for MAE/collateral/movement; rank and signed gaps use valid endpoints. Undefined rank values are not imputed or connected across missing points.

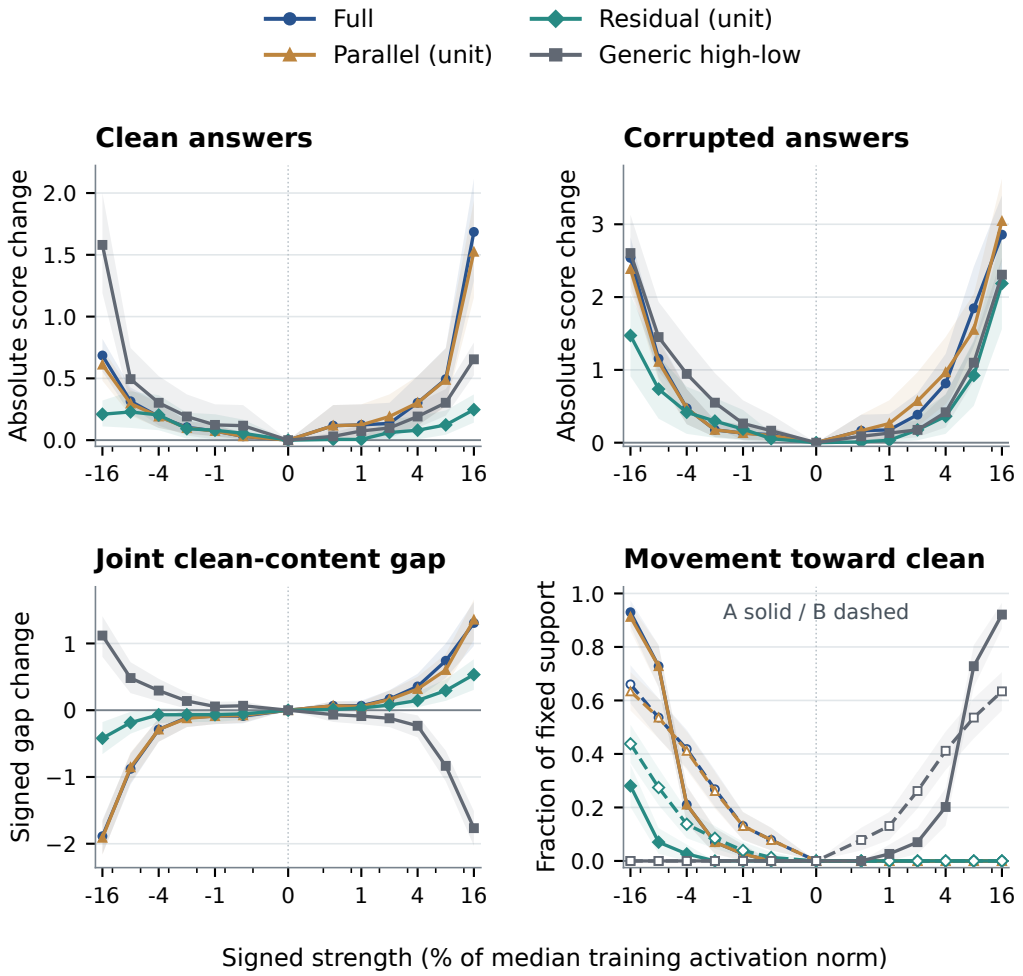


Figure 11: Bandwagon-negative collateral and movement curves under strict whole-response scoring. Absolute clean/content score changes quantify collateral. The signed joint-gap change compares jointly steered clean-minus-content scores with the original same-question gap. The final panel gives toward-clean fractions (A solid, B dashed) on fixed supports. All 13 signed strengths are shown on a symmetric-log x-axis (linear within 0.5%); colors/markers distinguish full, unit parallel, unit residual, and generic high–low directions in Llama block 23 using the expanded 11-dimensional reference. Shading is pointwise 95% benchmark-stratified question-cluster uncertainty, not simultaneous bands. No test strength is selected. Fixed supports: clean 162, A 114, B 153, content 91. Later invalid scores incur loss 9 for MAE/collateral/movement; rank and signed gaps use valid endpoints. Undefined rank values are not imputed or connected across missing points.

For Refinement, both strict framed supports contain 160 questions. Its strict response and collateral panels use the same direction signs, intervention strengths, and uncertainty procedure as the preceding primary-policy panels. Reading the two policies together distinguishes first-line rating changes from effects that also satisfy integer-only completion formatting, without choosing a separate intervention for the stricter endpoint.

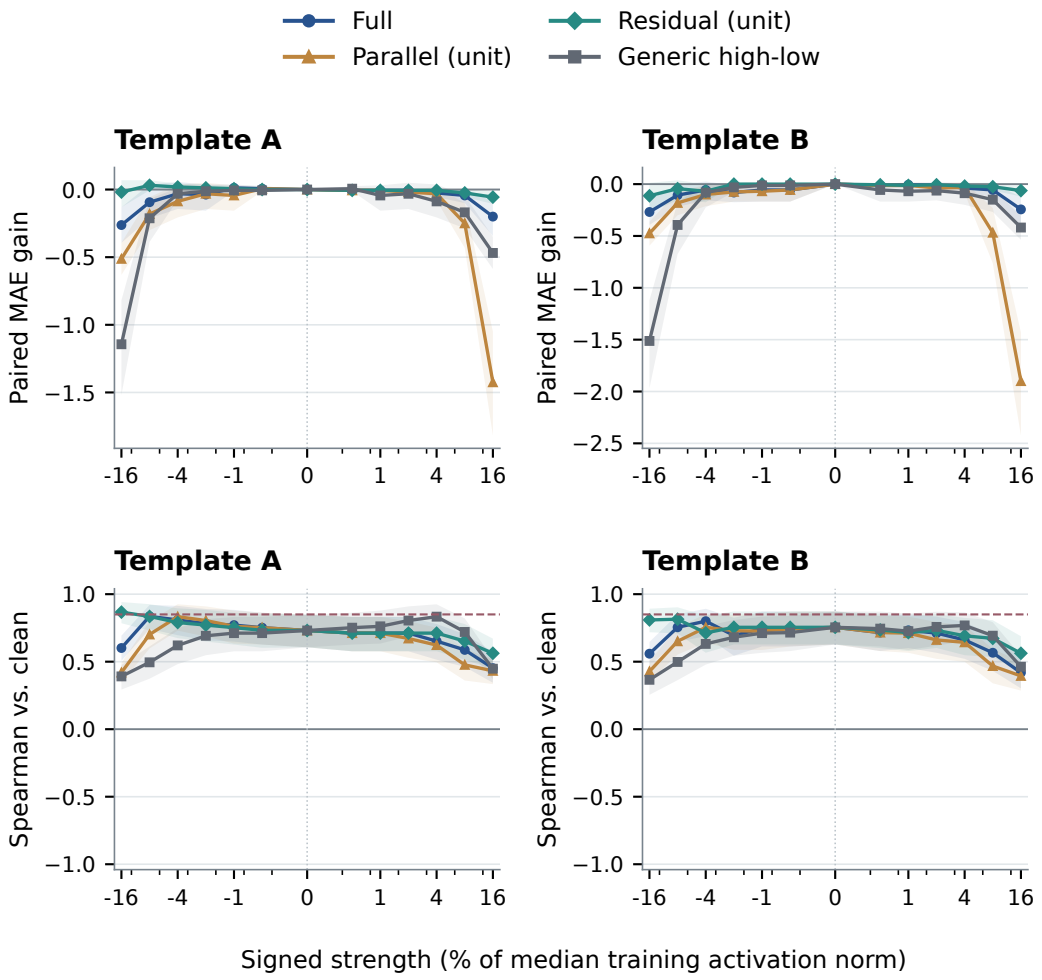


Figure 12: Refinement-negative response curves under strict whole-response scoring. Top: paired MAE gain toward each answer's clean score for templates A and B; positive is improvement. Bottom: clean-versus-steered Spearman; the dashed line marks the 0.85 ranking reference. All 13 signed strengths are shown on a symmetric-log x-axis (linear within 0.5%); colors/markers distinguish full, unit parallel, unit residual, and generic high–low directions in Llama block 23 using the expanded 11-dimensional reference. Shading is pointwise 95% benchmark-stratified question-cluster uncertainty, not simultaneous bands. No test strength is selected. Fixed supports: clean 162, A 160, B 160, content 91. Later invalid scores incur loss 9 for MAE/collateral/movement; rank and signed gaps use valid endpoints. Undefined rank values are not imputed or connected across missing points.

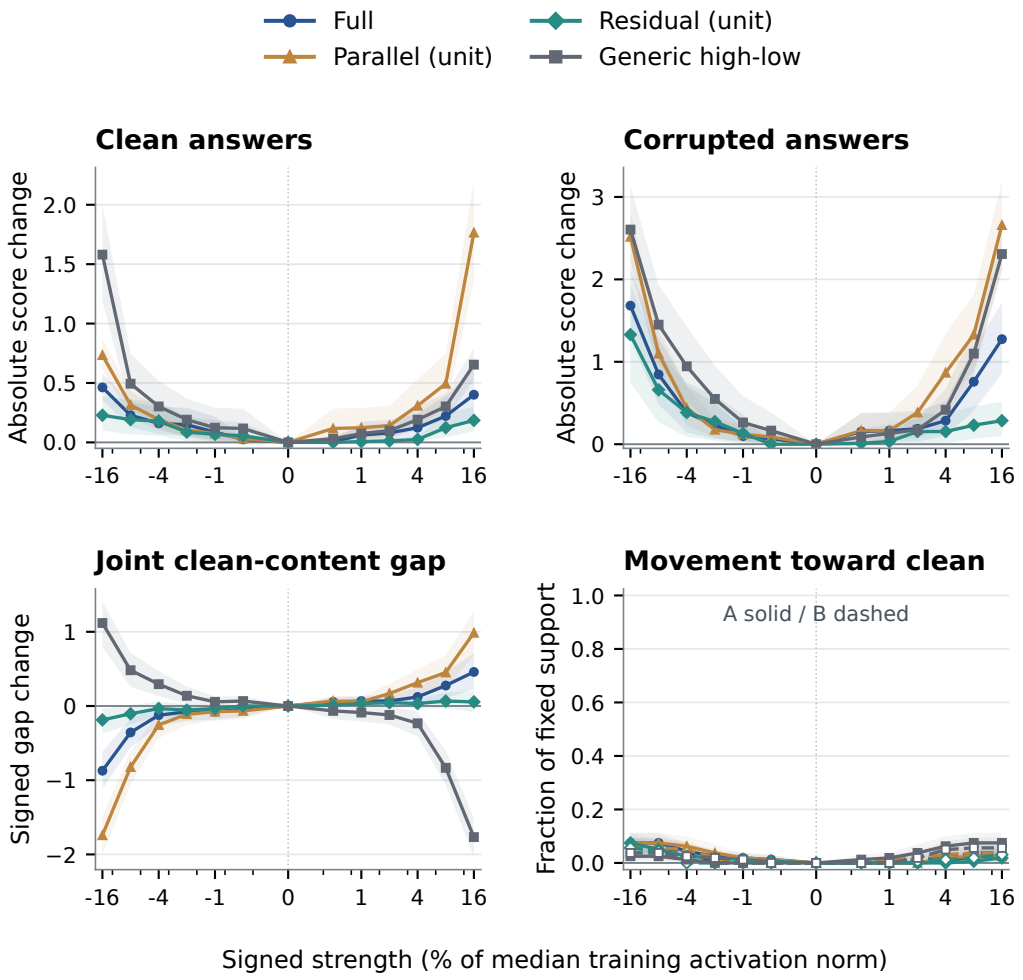


Figure 13: Refinement-negative collateral and movement curves under strict whole-response scoring. Absolute clean/content score changes quantify collateral. The signed joint-gap change compares jointly steered clean-minus-content scores with the original same-question gap. The final panel gives toward-clean fractions (A solid, B dashed) on fixed supports. All 13 signed strengths are shown on a symmetric-log x-axis (linear within 0.5%); colors/markers distinguish full, unit parallel, unit residual, and generic high–low directions in Llama block 23 using the expanded 11-dimensional reference. Shading is pointwise 95% benchmark-stratified question-cluster uncertainty, not simultaneous bands. No test strength is selected. Fixed supports: clean 162, A 160, B 160, content 91. Later invalid scores incur loss 9 for MAE/collateral/movement; rank and signed gaps use valid endpoints. Undefined rank values are not imputed or connected across missing points.

L LIMITATIONS

L.1 EVALUATION SETTING AND INPUT CONTROLS

The broad behavioral study covers seven judges, nine benchmarks, and seven framing families. The controlled study compares Llama-3.1-8B-Instruct and Qwen3-4B-Instruct-2507 on the same independently generated answers under a fixed scalar-rating prompt. Its three numerical and multiple-choice benchmarks support verifiable final-field changes, a different setting from pairwise preference evaluation, long-form writing, coding, or dialogue. The two metadata families concern consensus and review status, with one alternative wording per family in template B. Cross-model comparison evaluates the same relationships using directions fitted separately in each model; it does not test vector transfer between architectures. Framed/content intervention curves are evaluated in Llama.

Framing preserves the controlled study's answer body while changing the context available to the judge; the broad study also includes bounded prose and citation edits (Appendix B.1). Whether a response is undesirable depends on the evaluation policy, particularly when provenance or prior review counts as evidence. Here, metadata values are assigned experimentally. The neutral archive-code control measures a non-evaluative addition matched to its family's unfavorable-input token count; its effect can itself be nonzero. Final-answer corruption supplies a known error and may conflict with the retained reasoning; it represents this particular content change rather than naturally occurring quality variation in general. Human comparisons assess the broad study's sampled transformations, not the separate controlled cohort.

L.2 GEOMETRIC AND FUNCTIONAL INTERPRETATION

The clean-score spans provide comparison coordinates: scalar regression and a balanced high–low contrast define the common reference, and centered score-token logit coefficients expand the Llama reference. Orthogonality to these finite fitted spans does not imply absence of score information. Effective-pair geometry describes changes conditioned on a realized response; exact matching further selects a population with shared score-change strata. Spectra and mean-direction cosine characterize these populations rather than a unique causal pathway or equality of individual activation trajectories.

Interventions test additions at one block's final input-token position over a finite strength grid. Their effects characterize control at that position, not the full computation producing the original score. Raw component effects can be nonadditive, and unit components can receive different development-selected strengths. Equal mean score shifts at selected settings need not imply equal vectors, per-answer responses, or effects at other strengths. Forward and reverse constraints likewise have distinct targets: preserving clean-input ranking while moving ratings versus recovering ratings and rankings altered by framing.

L.3 MEASUREMENT AND UNCERTAINTY

Completed first-line ratings and strict whole-response compliance are distinct measurements. Strict-format sensitivity uses the primary-fitted configurations, with coverage that varies by model and domain. Common-reference comparisons and response curves fix the originally valid reference set; logit-augmented selected-configuration tables use the separate all-input loss convention in Appendix E.5. Their estimates must be interpreted with the corresponding population and loss rule. The scoring-policy adoption, refitting, and retrospective versus pre-acquisition protocol timing are detailed in Appendix D.4.

Question-cluster intervals preserve dependence among an answer's variants and comparisons, conditional on the collected cohort, fitted directions, exact matching, and development selection. They do not include retraining or model-selection uncertainty. Small matched sets and low-prevalence response groups are descriptive. The response curves show every sampled strength with pointwise intervals; the finite grid and fidelity requirement delimit the evaluated range rather than an unrestricted recovery frontier.