ONE CANDIDATE CHANGES THE REST: NON-LOCAL INSTABILITY IN LLM-BASED RANKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used to recommend and rank candidates in a variety of settings, including e-commerce and hiring. In these LLM-as-a-Recommender scenarios, models are often tasked with mapping an unordered set of candidates to an ordered ranking. While prior work has shown that LLM-based rankers can be sensitive to presentation order, we identify a distinct and previously underexplored property of these rankers: **Non-Local Instability**. Ideally, a non-locally stable ranker should ensure that modifying a single candidate affects only that candidate's position relative to the others. However, we show that current autoregressive LLMs systematically violate this property: small, localized changes to the description of a single candidate can alter the relative ordering of other candidates whose descriptions remain entirely unchanged. We characterize this phenomenon across multiple types of perturbations, evaluate its prevalence across different models, and investigate its underlying mechanisms. Our experiments show that non-local instability is widespread and persists even in recent state-of-the-art models. Reasoning can substantially mitigate this effect, but at the cost of increased ranking variability under the same context. These findings reveal a fundamental robustness and fairness concern in LLM-based ranking systems: the relative ranking assigned to a candidate can be substantially influenced by seemingly irrelevant semantic changes to other candidates.

1 INTRODUCTION

Large language models (LLMs) are increasingly used to compare candidates and support recommendations across domains. Consumers use generative AI for product research and recommendations (Pandya, 2025), while organizations are adopting AI tools for recruitment tasks such as screening resumes (SHRM, 2025). These applications can influence consequential decisions: a global survey found that approximately half of respondents had purchased products based on generative-AI recommendations (Capgemini Research Institute, 2025). Whether comparing products or evaluating applicants, ranking requires weighing multiple attributes against a user's needs. As LLMs take on this role, the reliability of their comparisons becomes increasingly important.

However, LLM's effectiveness does not necessarily imply such reliability. LLM-based rankers can achieve competitive performance on retrieval and recommendation benchmarks (Sun et al., 2023; Hou et al., 2024), and LLM relevance judgments can match human assessors in some search settings (Thomas et al., 2024). Yet ranking quality alone does not capture robustness: prior work documents sensitivity to candidate presentation order (Tang et al., 2024) and vulnerability to manipulations of candidate content (Pfrommer et al., 2024). Such sensitivity raises concerns about whether rankings reflect relevant evidence rather than incidental presentation choices, particularly when they influence access to opportunities or exposure.

Prior work on LLM-based ranking instability has primarily studied whether strategic modifications to a candidate's textual content can promote that candidate in the resulting ranking (Kumar & Lakkaraju, 2024; Pfrommer et al., 2024; Caville et al., 2026). Related work also finds that introducing an inferior alternative can bias hiring decisions (Valkanova & Yordanov, 2024). Our focus is on a different intervention: editing one existing candidate while holding the query, candidate set, presentation order, and all other descriptions fixed. We ask: *Can minor and localized changes to*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

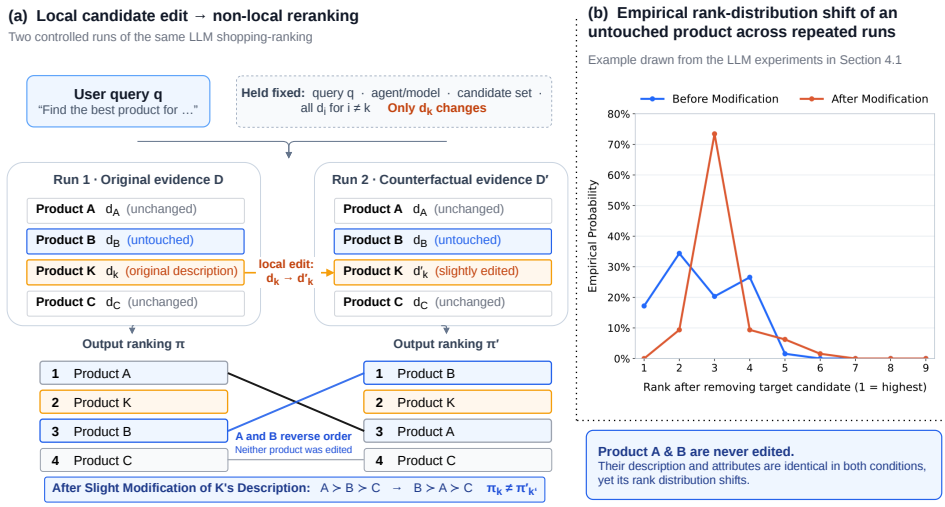


Figure 1: **Motivating examples showing non-local instability in LLM-based shopping ranking.** **(a)** A local edit to one product description can change the relative order of untouched products, even when the edited product's own rank remains unchanged. **(b)** An empirical example showing a shift in an untouched product's rank distribution after another candidate is edited. Ranks are computed after excluding the edited candidate.

one candidate's description or format affect not only its own ranking, but also the relative order of other candidate pairs?

We call this phenomenon ***non-local instability***. Figure 1 illustrates it through a shopping example: editing one product can reverse the order of two untouched products, even when the edited product's own position remains unchanged. This dependence also complicates description optimization: the apparent benefit of one candidate's edit can disappear when another competing candidate edits its description.

We demonstrate this phenomenon across various areas, including product recommendation, resume ranking, and controlled multi-attribute student ranking. Rather than equating a difference between two sampled rankings with an intervention effect, we compare rank distributions estimated from repeated samples. This separates the question of sensitivity to an edit from variability under an identical input, which single-run comparisons can conflate, as further discussed in App G.

We then investigate how local edits affect rankings among untouched candidates. Our experiments identify multi-attribute integration as a source of sensitivity: larger non-local shifts accompany larger changes in attribute preference, and directions extracted from minor edits can transfer behavioral preferences across candidate sets. Finally, we examine reasoning as a mitigation. Reasoning substantially reduces intervention sensitivity, but often increases variability across repeated rankings under identical inputs. And we show that clarifying the judging criteria largely mitigates this consistency cost, suggesting a practical solution.

Our main contributions are:

- **A distributional characterization of non-local instability in LLM-based ranking.** We measure changes in untouched candidates' rank distributions across different tasks and multiple models, distinguishing local-edit effects from finite-sampling variability.
- **Mechanistic evidence from controlled interventions.** We combine attribute-aggregation controls and representation interventions to trace and influence non-local effects.
- **An analysis of reasoning's benefits and limits on this phenomenon.** We distinguish robustness to local edits from within-condition ranking consistency, and proposed a trivial method to better help reasoning mitigate this instability.

2 Related Works

LLM-based Ranking. LLMs have been adapted to information retrieval and recommendation through listwise, pairwise, and setwise ranking formulations (Sun et al., 2023; Qin et al., 2024; Zhuang et al., 2024; Hou et al., 2024). Beyond ranking quality and efficiency, prior work examines sensitivity to candidate presentation order. Permutation self-consistency aggregates rankings across shuffled inputs (Tang et al., 2024), while InvariRank constrains cross-candidate attention and positional encoding to obtain position-invariant reranking (Bito et al., 2026). Another line of work studies whether strategic edits to candidate content can promote the edited candidate (Kumar & Lakkaraju, 2024; Pfrommer et al., 2024). Our focus is complementary: with the candidate set and input order fixed, we measure how a local edit affects the ranking distributions of candidates that were not edited, rather than only the target's promotion or robustness to permutations.

Bias in LLM-as-a-Judge. LLM judges exhibit position, verbosity, and self-enhancement biases (Zheng et al., 2023; Shi et al., 2025), as well as non-transitive pairwise preferences that can make aggregate rankings depend on the comparison baseline (Xu et al., 2025). Context effects are particularly relevant: in hiring decisions, an inferior alternative can change preferences between other candidates, and even irrelevant attributes of that alternative can amplify the effect (Valkanova & Yordanov, 2024). These findings establish that candidate evaluation can depend on the surrounding alternatives. We extend this perspective to repeated full-ranking outputs under localized edits, including edits that preserve all attributes. Removing the edited candidate before analysis isolates changes among untouched candidates from rank displacement caused solely by the target's movement.

Sampling Variability and Self-Consistency. LLM judges can produce inconsistent decisions across repeated runs with identical prompts and hyperparameters (Haldar & Hockenmaier, 2025). In reasoning tasks, self-consistency improves accuracy by aggregating answers from multiple sampled reasoning paths (Wang et al., 2023). More recently, Zollo et al. (2026) use empirical answer frequencies from offline repeated sampling to train a confidence predictor that requires only one generation at deployment. These studies examine or exploit variability under a fixed input, which is distinct from an intervention-induced change in the output distribution.

3 Preliminaries

For each evaluation case e, we modify one target candidate t among n candidates, keeping the query, input order, and all other descriptions fixed. We sample m complete rankings $\rho^a_{f,e,r}$ under each condition $a \in \{O, M\}$, where O denotes the original input, M the modified input, and f the model with its reasoning and decoding configuration.

Non-local instability. Non-local instability occurs when editing t changes an untouched candidate's rank distribution *after removing* t. Removing the target excludes numerical rank shifts caused solely by its movement. For $i = (e, j)$, $j \neq t$, let $p^a_{f,i}(k)$ denote the probability of target-removed rank k, estimated by its frequency $\widehat{p}^a_{f,i}(k)$ among the m samples. Candidate j is non-locally rank-stable when $p^O_{f,i} = p^M_{f,i}$. We calculate observed and calibrated total variation as

$$\widehat{D}_{f,i} = \frac{1}{2} \sum_{k=1}^{n-1} \left| \widehat{p}^O_{f,i}(k) - \widehat{p}^M_{f,i}(k) \right|, \qquad \widetilde{D}_{f,i} = \frac{[\widehat{D}_{f,i} - b_{f,i}]_+}{1 - b_{f,i}}, \tag{1}$$

where $[z]_+ = \max(z, 0)$ and $b_{f,i} < 1$ is the expected TV (called 'null TV') under random reassignment of O/M labels within case e. Across all evaluated case–candidate pairs $\mathcal{I}$, we report mean calibrated TV (cal. TV), $\overline{D}^{\text{cal}}_f = |\mathcal{I}|^{-1} \sum_{i \in \mathcal{I}} \widetilde{D}_{f,i}$, and the RMS-based non-local stability (NLS) score:

$$S^{\text{NL,cal}}_f = 1 - \sqrt{\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \widetilde{D}^2_{f,i}}. \tag{2}$$

Lower cal. TV and higher NLS indicate greater stability; we report NLS score on a 0–100 scale. Survival curves plot $C^{\text{cal}}_f(x) = \sum_{i \in \mathcal{I}} \mathbf{1}\{\widetilde{D}_{f,i} > x\}$, the number of non-target instances exceeding threshold x. Calibration details and metric interpretation are provided in Appendix A.

Within-condition consistency. This measures agreement between repeated rankings under the *same input and model configuration*. Let $d_K(\rho, \rho')$ be the fraction of candidate pairs ordered differently in two complete rankings. We calculate within-condition Kendall dispersion and its overall average as

$$K^a_{f,e} = \frac{1}{\binom{m}{2}} \sum_{r<s} d_K(\rho^a_{f,e,r}, \rho^a_{f,e,s}), \qquad \overline{K}_f = \frac{1}{2|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left(K^O_{f,e} + K^M_{f,e}\right), \tag{3}$$

where $\mathcal{E}$ is the set of evaluation cases. Lower dispersion indicates greater consistency.

Non-local instability measures *between-condition distribution shifts*; within-condition consistency measures *repeatability under an identical input*. Thus, smaller edit-induced shifts need not imply more consistent individual outputs.

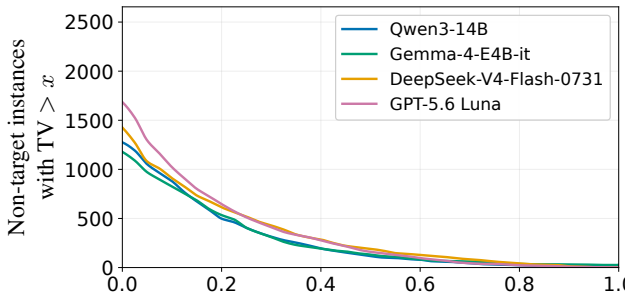
Datasets and Experimental Setup. Our data comprise a shopping pool of 120 queries spanning five product categories from the E-GEO test split (Bagga et al., 2025), five software-vacancy queries paired with de-identified resumes from the vacancy–resume matching dataset (Vanetik & Kogan, 2023), and synthetic student cohorts with randomly generated GPA, social-contribution, peer-evaluation, award-count, and teacher-score attributes. Each input contains ten candidates, with seeded candidate orders held fixed within each O/M comparison. We obtain repeated autoregressive outputs with decoding parameters fixed within each setting (temperature 0 for the main API panels), typically collecting 32–64 valid complete rankings per condition.Interventions affect only one candidate and include fact-preserving wording or name edits, added promotional cues, and controlled attribute changes. Experiment-specific sample sizes and decoding settings are reported with the corresponding analyses.

4 Local Edits Change Untouched Rankings

4.1 Non-Local Instability Phenomenon

Model	Mean Cal. TV↓	Cal. *N-L S*↑
Qwen3-14B	0.105	79.6
Gemma-4-E4B-it	0.104	78.8
DeepSeek-V4-Flash-0731	0.125	76.4
GPT-5.6 Luna	0.131	77.4

(a) Aggregate stability metrics.



(b) Calibrated total variation distance

Figure 2: **Local edits change untouched rankings.** All four models are evaluated on the same complete shopping and hiring comparisons without reasoning. **(a)** reports mean calibrated TV and calibrated RMS non-local stability score (NLS, 0–100). **(b):** At threshold x, the curve counts non-target instances with calibrated TV greater than x. Lower curves indicate smaller or less frequent non-local shifts. Curves are smoothed for display.

We first examine whether editing one candidate changes the rank distributions of candidates whose descriptions remain untouched. With reasoning disabled, we evaluate four models on matched shopping and hiring inputs, keeping the query, candidate set, and input order fixed within each original–modified comparison. Figure 2 summarizes 295 complete comparisons shared by all four models, yielding 2,655 non-target candidate instances per model. We remove the edited candidate before computing TV and calibrate against finite-sample variation. Thus, the measured shifts cannot be explained by that edited candidate moving up or down the list. Dataset composition, sampling configurations, and incomplete-input exclusions are documented in Appendix C.

Every model exhibits non-local shifts, including approximately 5% non-target instances per model with calibrated TV above 0.5. The survival curves distinguish their frequency from their magnitude:

height at small thresholds reflects how widespread the measured changes are, while the right tail reveals large shifts that a mean alone can obscure. Notably, without reasoning mode, the stronger API models do not consistently exhibit greater non-local stability than the smaller local models, suggesting that ranking stability does not automatically improve with general model capability. Task-specific curves are provided in Appendix C.

4.2 Attribute-Preserving Edits Are Sufficient

Non-local instability does not require changing a candidate's substantive attributes. In shopping, we replace the heading `Product attributes` with `Feature highlights`, capitalize ten selected words, or capitalize the full description and double its spaces. In hiring, we replace `Candidate details` with `Profile summary` while leaving the professional history unchanged. These interventions alter presentation without changing the supplied product specifications or qualifications. Table 1 reports calibrated NLS separately for each edit, making it possible to distinguish their effects from those of added promotional or endorsement cues.

Task	Edit	Qwen	Gemma	DeepSeek	GPT Luna
Shopping	Two-word heading	89.2	90.6	82.4	79.4
	Ten-word capitalization	92.1	84.2	82.3	79.5
	Full capitalization + spacing	62.9	62.3	67.9	68.3
	Sales-volume cue†	83.3	77.6	68.6	76.0
	Discount cue†	80.8	79.0	79.8	79.3
Hiring	Two-word heading	89.8	92.5	89.9	89.5
	Interviewer endorsement†	73.3	83.1	73.6	78.7

Table 1: **Non-local stability by intervention.** Calibrated RMS NLS (0–100); lower values indicate greater instability. Columns use the same four models as Figure 2. † marks added marketing or recommendation signals, not semantically invariant edits. Exact edits and coverage are specified in Appendix C.

The attribute-preserving interventions demonstrate that changing task-relevant evidence is not necessary to influence untouched candidates. Added sales, discount, and interviewer-recommendation cues provide a separate comparison: although only the target description changes, their effects can extend beyond that target. Appendix D further examines how edit extent relates to non-local effects through matched comparisons of smaller and larger edits.

4.3 Controlled Ranking of Synthetic Students

To separate presentation effects from ambiguity in real-world descriptions, we also rank synthetic cohorts of ten students using five explicitly supplied attributes: GPA, social-work contribution, peer evaluation, award count, and teacher score. Their ranges are $[1, 4]$, $[1, 10]$, $[1, 10]$, $\{0, \ldots, 10\}$, and $[1, 10]$, respectively. Our aligned student panel comprises 13 cohorts, two input orders, one selected target and three assigned edits per cohort–order combination, with 32 valid rankings per condition. Presentation edits leave all five attributes unchanged; details and task-separated results appear in Appendix C.4.

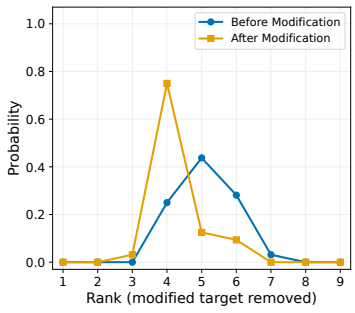


Figure 3: An untouched student's rank distribution shifts after another student is renamed. This is drawn from DeepSeek-v4-flash-0731 (w/o reasoning) experiments.

The example in Figure 3 illustrates the effect of changing only another student's name. Among the ten students, the untouched student shown ranks ninth in GPA, first in social contribution, tenth in peer evaluation, eighth in awards, and third in teacher score. Its sampled rank varies within both conditions, yet the distributions shift: after removing the edited student, the probability of fourth place rises from 25% to 75%, while that of fifth or sixth place falls from 71.9% to 21.9%. Its raw TV is 0.531, and calibrated TV

is 0.440. Thus, the effect concerns a change in the distribution, not merely disagreement between two sampled rankings. Even explicit numerical evidence and instructions to ignore names do not guarantee local effects. The full roster, exact name replacement, and distributions are provided in Appendix C.5.

5 Tracing and Controlling Non-Local Effects

5.1 Multi-Attribute Integration Is a Source of Sensitivity

Measuring attribute preference. To characterize how a ranking reflects its numerical evidence, we define *attribute alignment* as the expected fraction of student pairs ordered consistently with their scores on an attribute. Write $x_{i,d}$ for student i's score on attribute d, and let $\mathcal{P}_d$ contain pairs (i, j) with $x_{i,d} > x_{j,d}$. Then

$$a_d = \frac{1}{|\mathcal{P}_d|} \sum_{(i,j)\in\mathcal{P}_d} \Pr(i \text{ precedes } j).$$

Higher alignment means the output more closely follows that attribute's descending score order; tied scores are excluded. For local-edit comparisons, we also exclude the edited student and summarize the changes across the five attributes as $\Delta_{\text{align}} = [\frac{1}{5}\sum_d (a_d^M - a_d^O)^2]^{1/2}$. The higher Δ_{align} indicates a larger shift of the model's preference on different attributes.

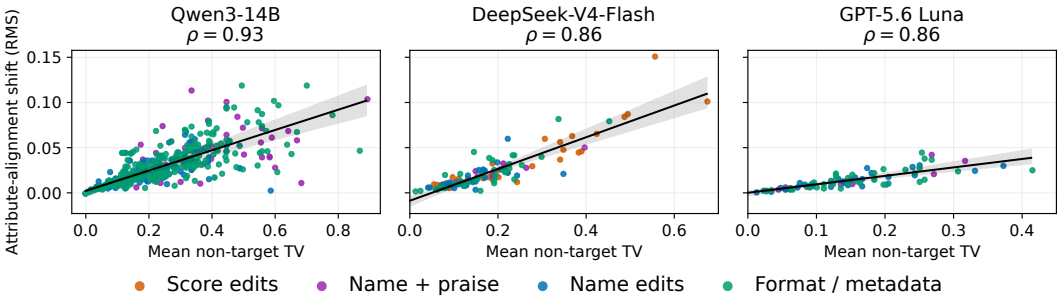


Figure 4: **Non-local shifts accompany changes in attribute preference.** Each point is an original/modified comparison in a fixed five-attribute student task. The horizontal axis is mean target-removed rank TV; the vertical axis is the RMS change in pairwise agreement with the five attribute orderings. Colors distinguish edit families. Lines are descriptive linear fits with profile-cluster bootstrap bands; panels are fitted separately because model coverage differs. Coverage and definitions appear in Appendix E.

Larger distribution shifts accompany larger changes in attribute alignment. Figure 4 combines archived name, praise, formatting, metadata, and score interventions: larger non-local distribution shifts tend to accompany larger changes in attribute preference (Δ_{align}). This association is consistent with edits affecting the effective trade-offs among attributes, even though the remaining students' scores are unchanged.

Multi-attribute ranking is more sensitive than a single-criterion task. We vary the available attributes while fixing cohorts, candidate orders, targets, edits, and sampling counts. Starting with GPA alone, we successively add the other four original attributes and four synthetic attributes (Appendix E). In Figure 5, Model's calibrated NLS falls from 94.2 with GPA alone to 83.4 with two attributes and 80.6 with five. Mean raw TV rises from 0.010 to 0.124 and 0.153, respectively; its null expectation also increases, but the difference remains after calibration. These results support multi-attribute integration as a source of sensitivity.

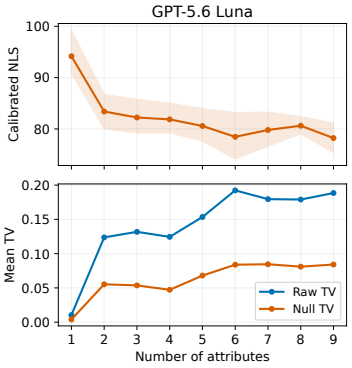
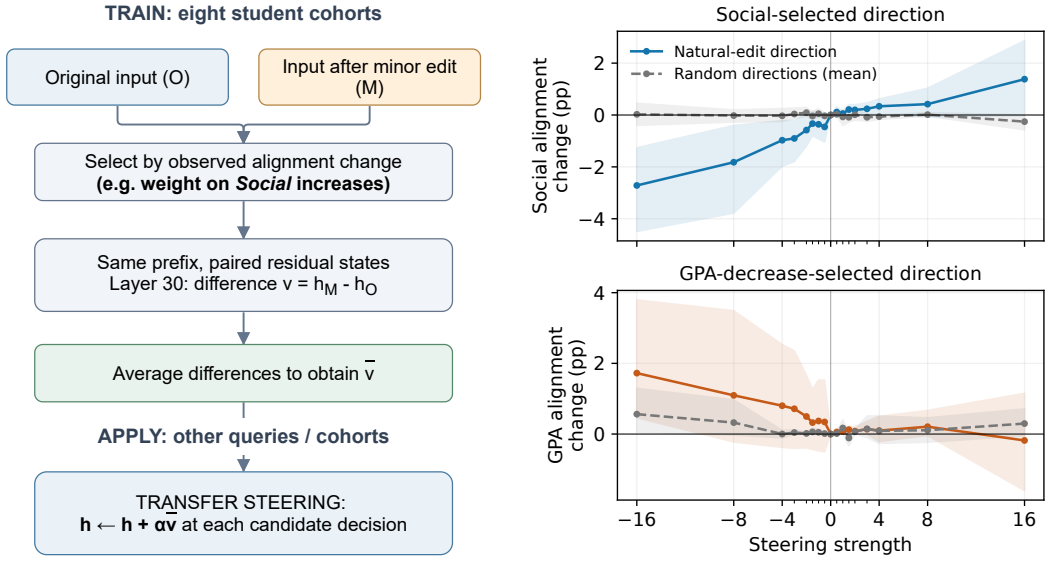


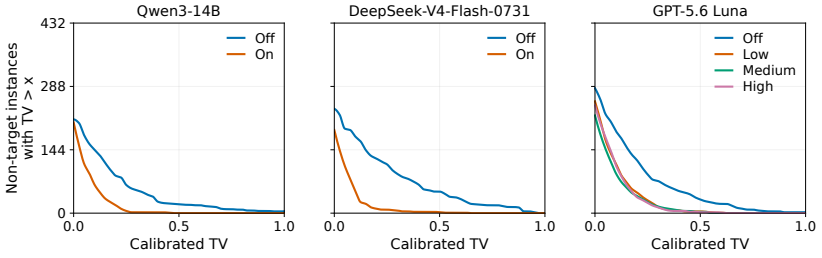
Figure 5: Attribute-count intervention on GPT-5.6 Luna.



(a) Extracting directions from natural edits.

(b) Dose response on held-out cohorts.

Figure 6: **Transferring edit-derived directions without editing the test input.** **(a)** Attribute-preserving edits are selected by their observed social or GPA alignment change; paired residual differences are averaged into fixed directions. **(b)** demonstrates changes in the selected attribute's preference after steering against w/o steering, across five held-out cohorts and two candidate orders. Signed strength scales the mean residual difference. Gray curves average three norm-matched random directions. Bands are descriptive cohort-bootstrap 95% intervals. All five attribute responses are reported in Appendix F.



Calibrated NLS ↑

Model	Mode	Score
Qwen	Off	78.32
	On	92.35
DeepSeek	Off	71.88
	On	92.25
Luna	Off	75.48
	Low	88.10
	Medium	88.77
	High	88.65

Figure 7: **Reasoning reduces sensitivity to local interventions.** The first three panels compare reasoning settings on identical shopping cases. At threshold x, each curve counts non-target instances with calibrated TV greater than x, after removing the edited candidate; lower curves indicate fewer shifts exceeding that magnitude. Curves use monotone interpolation for display. The rightmost table reports calibrated RMS NLS (0–100; higher is better), computed from unsmoothed values. Each configuration contains 432 instances from 48 queries, not 432 independent queries.

5.2 Local Edits Yield Transferable Behavioral Steering Directions

From local edits to fixed directions. We ask whether behavior-associated components of local edits can transfer to other candidate sets. Using Qwen3-14B, we select numerical-attribute-preserving edits that increase social alignment or decrease GPA alignment on eight training cohorts. For each selected edit, we compute differences at layer index 30 under identical generation prefixes, average within each cohort, and then average across cohorts. This yields one fixed direction per selection criterion, without explicit attribute-priority instructions (Rimsky et al., 2023). On five held-out cohorts, we leave the input unchanged and add the scaled direction at each candidate-label decision during natural sampled generation. (Appendix F).

Transfer is directional but asymmetric. Figure 6 shows a signed dose response for the social-selected direction: social alignment changes by -2.72 percentage points at strength -16 and $+1.38$ at $+16$, compared with the unsteered baseline. Transfer is weaker for the GPA-decrease-selected direction. One potential reason is that increasing a certain attribute's alignment can be more common and easier than solely decreasing one. Thus, edit-derived residual differences contain components that can shift ranking behavior on new candidate sets, but they are not symmetric or independent attribute-weight controls, further analysis can be seen in Appendix F. These interventions establish partial causal controllability. The frozen directions also transfer to some differently formatted inputs, including a social-alignment decrease under whole-roster JSON serialization, although effects depend on the direction, intervention sign, and format (Appendix F.1, Figure 13).

6 Reasoning's Influence on Non-Local Stability

6.1 Reasoning Reduces Intervention Sensitivity

We compare reasoning settings on a matched shopping subset of 48 queries. Each original and modified input has 32 valid sampled rankings, yielding 432 target-removed candidate instances per configuration. We evaluate Qwen3-14B and DeepSeek-V4-Flash-0731 with reasoning off and on, and GPT-5.6 Luna at off, low, medium, and high effort. The subset covers the five shopping edit types used above.

Figure 7 shows substantially smaller non-local shifts with reasoning enabled. The cal. TV survival curves show that this improvement is not merely a change in the aggregate score: reasoning also reduces the prevalence of larger distribution shifts. However, residual sensitivity remains, and increasing effort beyond low does not produce a monotonic improvement. These results support reasoning as a mitigation on the matched inputs, not a guarantee that untouched candidates' ranking distributions remain invariant.

6.2 Intervention Stability versus Sampling Consistency

What accompanies this reduction in intervention sensitivity? On the same queries, reasoning produces less consistent rankings across repeated samples of an identical input. Table 2 reports normalized Kendall dispersion over complete rankings, calculated within each original or modified condition before averaging, as defined in Equation 3. GPT-5.6 Luna's overall dispersion increases from 0.0623 with reasoning off to over 0.09 with reasoning enabled. Thus, the original and modified distributions can become closer while each becomes more dispersed: robustness to an intervention and repeatability under a fixed input are distinct properties. Per-model rank-distribution examples appear in Appendix G.

Table 2: **Within-condition Kendall dispersion** on the same 48 shopping queries. Lower is more consistent.

Model	Reasoning	Dispersion ↓
Qwen3-14B	Off	0.0455
	On	0.1424
DeepSeek-v4-flash-0731	Off	0.0477
	On	0.0960
GPT-5.6 Luna	Off	0.0623
	Low	0.0945
	Medium	0.0983
	High	0.0953

6.3 Clarifying Judging Criteria Mitigates the Consistency Cost

The sampling-consistency cost of reasoning can be largely mitigated by a simple clarification of the judging criteria. We test Qwen3-14B on synthetic student rankings, crossing reasoning off/on with holistic evaluation or an explicit rule: normalize the five attributes to $[0, 1]$, average them with equal weights, and break ties by student ID. Four cohorts, one candidate order, and two attribute-preserving edits are held fixed across these settings, with 16 samples per input.

Setting	Cal. NLS ↑	Kendall dispersion ↓
Neither	51.83	.0185
Criteria only	62.24	.0156
Reasoning only	87.52	.1649
Criteria + reasoning	**96.42**	**.0042**

Table 3: **Simple criteria clarification mitigates the consistency cost.**

Table 3 shows that reasoning alone reduces non-local sensitivity but increases within-condition Kendall dispersion from 0.0185 to 0.1649. Adding the explicit criteria reduces this dispersion to 0.0042, a 97.5% reduction relative to reasoning alone, while further improving calibrated Non-Local Stability Score from 87.52 to 96.42. A complementary DeepSeek student experiment with 32 samples per input likewise observes zero non-local TV and zero within-condition dispersion under the complete rule with reasoning, with every sampled ranking following that rule. Thus, in these controlled tasks, clarifying how evidence should be combined allows reasoning to improve both intervention stability and sampling consistency. Full setups and results are in Appendix H.

This mitigation should not be interpreted as a guarantee that reasoning combined with stated user preferences eliminates non-local instability in complex and realistic scenrios. Appendix H.1 illustrates a substantial empirical rank-distribution shift even at high reasoning effort in a shopping query with rather clear criteria.

7 Discussion

Description optimization depends on other candidates. Non-local instability makes it difficult for candidates to assess and maintain the benefit of improving their own descriptions. A merchant may find an edit beneficial against a fixed catalog, yet cannot assume that the gain persists when competitors also revise their listings. To illustrate this dependence, we append `[OVER 20K+ SOLD!]` first to product A and then to B, leaving C unchanged. Table 4 shows that editing A alone raises $P(A > C)$ from 0% to 100%. When B receives the same cue, $P(A > C)$ returns to 0%, despite no further change to either A or C. Thus, the apparent gain disappears under a change outside the merchant's control. This may make isolated optimization an unreliable guide to future benefit or harm.

Condition	$P(A > C)$	$P(A > B)$	$P(B > C)$
Original	0%	0%	100%
Edit A	100%	18.75%	81.25%
Edit $A + B$	0%	0%	100%

Table 4: **A local gain can disappear after a competitor edits its description.** 64 samples per condition (w/o reasoning). C is unchanged.

Potential implications for fairness. Candidates should not have control over the judging criteria. Our results raise a concern beyond direct self-promotion: a strategically written entry (which can be totally allowed, unlike prompt injections) may alter the effective trade-offs used to compare other candidates, favoring some attribute profiles over others without changing those candidates' qualifications. In recruitment or student selection, this would make evaluation depend partly on another participant's presentation choices rather than solely on the stated criteria and the candidates' records. The association between non-local shifts and attribute alignment, together with the transfer of edit-derived steering directions (Section 5.2), motivates this concern but does not establish a reliable text-only attack, complete control of attribute weights, or measured demographic discrimination. Fairness evaluations should therefore examine whether an untouched candidate's treatment remains stable under changes to other entries, alongside conventional accuracy and group-level measures. Explicit judging rules can reduce this sensitivity, but the residual shopping effects caution against treating reasoning alone as a fairness safeguard.

8 Conclusion, Limitations, and Future Work

We reveal non-local instability in LLM-based ranking across various domains: editing one candidate can change the rank distributions of untouched candidates. Our experiments implicate multi-attribute integration as a source of sensitivity and demonstrate partial causal control through transferable, minor-edit-derived steering directions. Reasoning reduces intervention sensitivity but can increase within-condition variability; clarifying judging criteria largely mitigates this cost in controlled student tasks. These findings are limited to the evaluated models, domains, and ten-candidate settings, and do not establish a complete mechanism or universal remedy. Finite-sample calibrated rank marginals also do not capture all joint-ranking changes, and stability alone does not guarantee quality or fairness. Future work should examine how these effects propagate through real agents and multi-stage systems, extend evaluation to larger and evolving candidate pools, and develop mitigations that preserve ranking quality.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the generation of experimental code. We did not use generative AI tools for other tasks requiring disclosure, and the remaining required-disclosure categories are not applicable to this work. Additionally, we used generative AI tools for language polishing to improve the grammar, clarity, and readability of the manuscript. All AI-generated code was manually reviewed, tested, and validated by the authors before being used in our experiments. All AI-assisted edits to the manuscript were also reviewed by the authors to ensure that they accurately reflected the intended scientific content and claims. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main text and appendices document our datasets, perturbations, sampling settings, evaluation metrics, and intervention procedures. Our anonymous repository provides analysis code, archived numerical results, and sampling and steering implementations: https://anonymous.4open.science/r/nonlocal-ranking-reproducibility/. The accompanying README describes the code structure, dependencies, and commands for recomputing the main metrics and figures.

REFERENCES

Puneet S. Bagga, Vivek F. Farias, Tamar Korkotashvili, Tianyi Peng, and Yuhang Wu. E-GEO: A testbed for generative engine optimization in e-commerce. *arXiv preprint arXiv:2511.20867*, 2025.

Ethan Bito, Yongli Ren, and Estrid He. One pass, any order: Position-invariant listwise reranking for LLM-based recommendation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 2026. doi: 10.1145/3805712.3809952.

Capgemini Research Institute. From hype to habit: How consumers are embracing ai. Technical report, Capgemini Research Institute, 2025. URL https://www.capgemini.com/wp-content/uploads/2025/12/Research-brief_From-hype-to-habit.pdf.

Evan Caville, Spencer Kayser, Siamak Layeghy, Billy Sung, Sara Dolnicar, and Marius Portmann. SIREN (luring LLMs onto the rocks): PAIR-driven preference manipulation in web-RAG recommenders, 2026. URL https://arxiv.org/abs/2607.21951.

Rajarshi Haldar and Julia Hockenmaier. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24986–25004. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1361. URL https://aclanthology.org/2025.findings-emnlp.1361/.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *Advances in Information Retrieval*, volume 14609 of *Lecture Notes in Computer Science*, pp. 364–381. Springer, 2024. doi: 10.1007/978-3-031-56060-6_24.

Aounon Kumar and Himabindu Lakkaraju. Manipulating large language models to increase product visibility, 2024. URL https://arxiv.org/abs/2404.07981.

Vivek Pandya. Generative ai-powered shopping rises with traffic to u.s. retail sites. Adobe Digital Insights, August 2025. URL https://business.adobe.com/blog/generative-ai-powered-shopping-rises-with-traffic-to-retail-sites.

Samuel Pfrommer, Yatong Bai, Tanmay Gautam, and Somayeh Sojoudi. Ranking manipulation for conversational search engines. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9523–9552. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.534.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1504–1518. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.97.

Nina Rimsky, Nick Gabrieli, Julia Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *ArXiv*, abs/2312.06681, 2023. URL https://api.semanticscholar.org/CorpusID:266174252.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025.

SHRM. Candidate "ghosting" and employer competition are fueling talent shortages, new SHRM research finds. Press release accompanying the 2025 Talent Trends report, July 2025. URL https://www.shrm.org/about/press-room/candidate--ghosting--and-employer-competition-are-fueling-talent. Published July 25, 2025; accessed September 24, 2026.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14918–14937. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.923.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Türe. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2327–2340. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.129.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1930–1940. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657707.

Kremena Valkanova and Pencho Yordanov. Irrelevant alternatives bias large language model hiring decisions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6899–6912. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.405.

Natalia Vanetik and Genady Kogan. Job vacancy ranking with sentence embeddings, keywords, and named entities. *Information*, 14(8):468, 2023. doi: 10.3390/info14080468. URL https://www.mdpi.com/2078-2489/14/8/468.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. Investigating non-transitivity in LLM-as-a-judge. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69583–69612. PMLR, 2025.

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*, volume 36, pp. 46595–46623, 2023.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 38–47. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657813.

Thomas Zollo, Jimmy Wang, and Richard Zemel. Unsupervised confidence calibration for reasoning LLMs from a single generation. *arXiv preprint arXiv:2604.19444*, 2026. doi: 10.48550/arXiv.2604.19444. URL https://arxiv.org/abs/2604.19444.

A Metric Computation and Interpretation

Target-removed rank distributions. For a complete ranking ρ, let $\rho \setminus t$ denote the ranking obtained by deleting the target and reindexing the remaining candidates from 1 to $n-1$. For example, (A, t, B) and (t, A, B) both reduce to (A, B). For $i = (e, j)$, $j \neq t$, the empirical rank probabilities used in Equation 1 are

$$\widehat{p}^{\,a}_{f,i}(k) = \frac{1}{m}\sum_{r=1}^{m} \mathbf{1}\Big\{\mathrm{rank}_{\rho^a_{f,e,r}\setminus t}(j) = k\Big\}, \qquad k \in \{1, \ldots, n-1\}. \tag{4}$$

Each distribution is estimated separately for a fixed query, candidate order, target, edit, and condition. The TV distance lies in $[0, 1]$: zero indicates identical empirical rank distributions, and one indicates disjoint rank supports. These are candidate-level *marginal* distributions; their equality does not generally imply equality of the joint distribution over complete rankings.

Finite-sample calibration. Even identical underlying distributions can yield positive empirical TV. To obtain an item-specific null baseline, we pool the $2m$ rankings from O and M within a case and uniformly reassign them to two groups of size m. Entire rankings are reassigned together, preserving dependencies among candidates. No rankings from different cases are pooled. Let $N_{f,i}(k)$ be candidate j's pooled count at target-removed rank k. Its count in one reassigned group follows $X_{f,i}(k) \sim \mathrm{Hypergeom}(2m, N_{f,i}(k), m)$. By linearity of expectation, the exact conditional null expectation is

$$b_{f,i} = \frac{1}{2}\sum_{k=1}^{n-1} \mathbb{E}\left[\left|\frac{X_{f,i}(k)}{m} - \frac{N_{f,i}(k) - X_{f,i}(k)}{m}\right|\right]. \tag{5}$$

This label-permutation baseline is not a within-condition dispersion measure. The calibrated TV in Equation 1 subtracts the expected null distance, clips negative values to zero, and rescales by $1 - b_{f,i}$. It is a normalized excess-distance statistic, not an unbiased estimate of population TV or a significance test. Zero calibrated TV means the observed distance does not exceed its baseline, not that the underlying distributions are proven equal. Cases with $b_{f,i} = 1$ have an undefined calibrated score and require additional samples.

Aggregate non-local metrics. The RMS in Equation 2 emphasizes large shifts more than an arithmetic mean. A reported NLS of 100 means every observed distance is at or below its null baseline; it does not establish exact invariance or deterministic outputs. The corresponding survival-count function is

$$C^{\mathrm{cal}}_f(x) = \sum_{i \in \mathcal{I}} \mathbf{1}\Big\{\widetilde{D}_{f,i} > x\Big\}, \qquad x \in [0, 1]. \tag{6}$$

Each counted instance is a case–candidate pair, so a candidate evaluated under different inputs contributes separately.

Within-condition Kendall dispersion. For complete rankings without ties, the normalized Kendall distance is

$$d_K(\rho, \rho') = \frac{1}{\binom{n}{2}} \sum_{u<v} \mathbf{1}\left\{\left(\text{rank}_\rho(u) - \text{rank}_\rho(v)\right)\left(\text{rank}_{\rho'}(u) - \text{rank}_{\rho'}(v)\right) < 0\right\}, \tag{7}$$

where $u < v$ enumerates unordered candidate pairs using any fixed indexing. Equation 3 averages over all distinct pairs of sampled rankings *within each condition*, then averages the resulting dispersions over O, M, and evaluation cases. It estimates the expected fraction of pairwise preferences that disagree between two independent outputs under identical inputs. For example, a dispersion of 0.10 means that 10% of candidate-pair preferences disagree on average. Zero empirical dispersion means all sampled rankings in that condition are identical. When reporting the target-removed variant, we apply the same calculation to $\rho \setminus t$ and normalize by $\binom{n-1}{2}$.

Within-condition dispersion and non-local TV answer different questions. A model can produce diverse rankings under a fixed input yet have identical rank distributions under O and M; conversely, it can always produce one target-removed ranking under O and a different target-removed ranking under M, giving zero within-condition dispersion but nonzero non-local TV. Consequently, an improvement in one metric does not by itself imply an improvement in the other.

B Sampling Strategy

Why repeated stochastic sampling? Our estimand is the change in an untouched candidate's rank distribution, rather than disagreement between two individual outputs. A single pair of rankings cannot separate an intervention effect from within-input variability, while repeating greedy decoding does not sample the stochastic output distribution. We therefore collect multiple independently requested generations for every fixed O/M input, estimate rank probabilities separately within each condition, and calibrate empirical TV against finite-sample variation as described in Appendix A.

Evaluation configuration. The main-text local experiments use stochastic sampling, not `do_sample=False`. The matched shopping–hiring comparison uses unquantized BF16 weights, `do_sample=True`, temperature 0.7, top-$p = 0.8$, top-$k = 20$, and a 256-token output cap with reasoning off. Each condition contains 32 or 64 valid complete rankings according to its panel, as specified in Appendix C. Query text, candidate order and non-target entries remain fixed across O/M. Repeated API calls use the model-specific configurations in Appendix C.6; an API temperature setting alone is not a guarantee of deterministic outputs. Invalid or truncated outputs are replaced, and comparisons with an unfilled valid-sample quota are excluded from the matched analysis.

Greedy comparison on matched inputs. To illustrate sensitivity to the generation configuration, we also report a separate Qwen3-14B greedy comparison on eight shopping queries, two fixed orders, ten candidates, one target, and the heading edit from `Product attributes:` to `Feature highlights:`. Each configuration has 64 outputs per O/M condition and $8 \times 2 \times 9 = 144$ non-target instances. The greedy arm explicitly uses `do_sample=False`; its repeated calls are not independent draws from stochastic decoding. Table 5 and Figure 8 compare their empirical metrics with stochastic sampling. Both use the same BF16 Qwen checkpoint and Transformers 4.57.6, but PyTorch differs (2.13.0 versus 2.7.0) and the output caps are 512 versus 256 tokens. This is therefore a configuration comparison, not a decoding-only causal ablation. Greedy results are confined to explicitly identified appendix comparisons.

Non-local effects persist under sampling. The raw-TV 90th percentile is 0.734 under the greedy configuration and 0.281 under stochastic sampling; the corresponding numbers of instances above 0.5 are 17 and zero. Nevertheless, stochastic sampling yields mean calibrated TV 0.061 and calibrated NLS 88.8. We test equality of the two rank distributions for each of its 144 instances using 200,000 pooled-label permutations, the raw-TV statistic, and the correction $(b+1)/(B+1)$, where b counts permuted statistics at least as large as observed. Benjamini–Yekutieli correction at a 5% false-discovery rate retains 13 instances (9.0%). These are dependent candidate instances, not independent queries or a population prevalence estimate. No analogous inferential claim is made from repeated greedy calls.

Qwen3-14B	Greedy decoding	Stochastic sampling
Reasoning	Off	Off
Weight precision	BF16	BF16
`do_sample`	False	True
Temperature	0 (inactive)	0.7
Top-p / top-k	Inactive	0.8 / 20
Output-token cap	512	256
Queries $\times$ orders	8×2	8×2
Outputs per O/M condition	64 repeated calls	64 random samples
Non-target TV instances	144	144
Mean raw TV	0.121	0.098
Raw TV p90	0.734	0.281
Raw TV > 0.5	17	0
Mean calibrated TV	0.116	0.061
Calibrated NLS	69.1	88.8

Table 5: Greedy decoding and stochastic sampling on the same 144 Qwen3-14B instances. Calibrated NLS is on a 0–100 scale. Greedy metrics are descriptive empirical summaries, not inferential evidence from independent random samples.

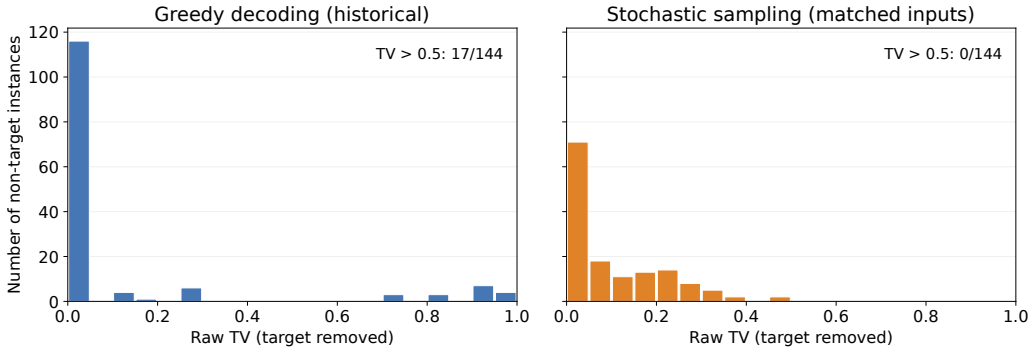


Figure 8: Empirical raw target-removed TV distributions on matched inputs. These are frequency histograms, not survival curves. Both panels use bins of width 0.05 and shared axes; no smoothing is applied, and the last bin includes TV equal to one. Each histogram contains 144 instances.

For example, an untouched candidate is second in all 64 Original samples but first in 29/64 Modified samples after target removal. Raw TV is 0.453, expected null TV is 0.060, and calibrated TV is 0.418. Its untouched competitor moves from first in every Original sample to second in 29/64 Modified samples. On the larger matched shopping–hiring panel, stochastic Qwen and Gemma have 117 and 132 of 2,655 instances with calibrated TV above 0.5, respectively. These latter threshold counts describe effect sizes, not significance-test discoveries.

C Phenomenon Results by Domain

Matched coverage and aggregation. The main comparison in Figure 2 includes shopping and hiring, with reasoning disabled. We retain an original–modified comparison only when all four models have the prescribed number of valid outputs under both conditions. The same query, candidate order, target, edit, and nine non-target candidates therefore contribute to every model's curve. Table 6 distinguishes the planned comparisons from this common complete subset. Original outputs are reused across edits where applicable. A repeated query or a second non-target candidate does not constitute an independent query observation. Students are analyzed separately rather than being mixed into the main shopping–hiring model comparison.

Panel	Planned pairs	Retained pairs	TV instances
Shopping: 32 samples	240	227	2043
Shopping: 64 samples	16	16	144
Hiring	60	52	468
Total	316	295	2655

Table 6: Coverage of the matched four-model comparison. One O/M comparison contributes nine non-target TV instances. The 32-sample and 64-sample shopping panels have different designs and may overlap in query identity; their query counts should not be added as disjoint sets.

Output failures and exclusions. We require a complete permutation containing each supplied candidate ID exactly once. Whitespace and letter-case normalization are allowed; invented IDs, missing or repeated candidates, single-character substitutions, truncated outputs, and unexpected reasoning in an off condition are invalid. Invalid attempts are retained and do not invalidate a comparison if replacement samples complete its prescribed count. Inputs that remain below their prescribed valid count after the retry budget are excluded. Table 7 reports both unsuccessful attempts and unresolved conditions; these are different quantities. The common-support analysis excludes a comparison for *all* four models if it is incomplete for either local model. These exclusions select for successful format compliance and can limit generalization; the resulting rank distributions are conditional on valid output, not distributions that include a failure outcome.

Model	Total attempts	Invalid attempts	Failure rate	Incomplete inputs	Missing outputs	Excluded pairs
Qwen3-14B	24,817	6,513	26.24%	2/467	64/18,368	2/316
Gemma-4-E4B-it	62,770	42,061	67.01%	22/571	987/21,696	19/316

Table 7: Local sampling failures and denominators. Attempt failure rate is invalid attempts divided by total attempts, including retries. The remaining columns give incomplete/total O or M inputs, missing/planned valid outputs, and excluded/planned shopping–hiring O/M comparisons before intersection across models. Qwen's collection covers shopping/hiring; Gemma's also includes the completed student panel.

Sampling, calibration, and smoothing. The local runs use unquantized BF16 weights, stochastic sampling with temperature 0.7, top-$p = 0.8$, top-$k = 20$, and a 256-token output cap with thinking disabled. API models use their model-specific provider and decoding configurations; matched inputs do not imply identical samplers across architectures and providers (Appendix C.6). We recompute every TV and its exact pooled-label permutation expectation from the saved counts, using Equations 1 and 5. We aggregate item-level calibrated distances, never rankings from different inputs. Survival curves use monotone cubic interpolation on a threshold grid of width 0.025, preserving point masses at zero and one. Metrics use unsmoothed distances. A positive calibrated distance is an excess-distance estimate, not by itself a significance-test discovery.

C.1 SHOPPING

The 32-sample E-GEO panel contains 120 queries spanning Arts, Crafts & Sewing; Industrial & Scientific; Office Products; Sports & Outdoors; and Tools & Home Improvement. Each query has ten candidate products, one seeded input order, one target, and two assigned edits from five types, with 32 valid samples per condition. This is not a full query-by-edit factorial. The 64-sample panel contains eight queries, two input orders, one target, the two-word heading edit, and 64 valid samples per condition. The shopping curves use the common complete subset from both panels.

The five edits are: `Product attributes:` to `Feature highlights:`; capitalizing ten selected words; capitalizing the target's full description and doubling spaces; appending `[OVER 20K+ SOLD!]`; and appending `[ON SALE! ALREADY 10% DIESCOUNT!]`. The last phrase retains the experimental spelling. The first three preserve supplied attributes, whereas the last

two add promotional claims. Token-count preservation is tokenizer-specific; the frozen text is not changed separately for each model merely to force matching token counts. Rows in Table 1 match inputs across models, but different edits can involve different queries.

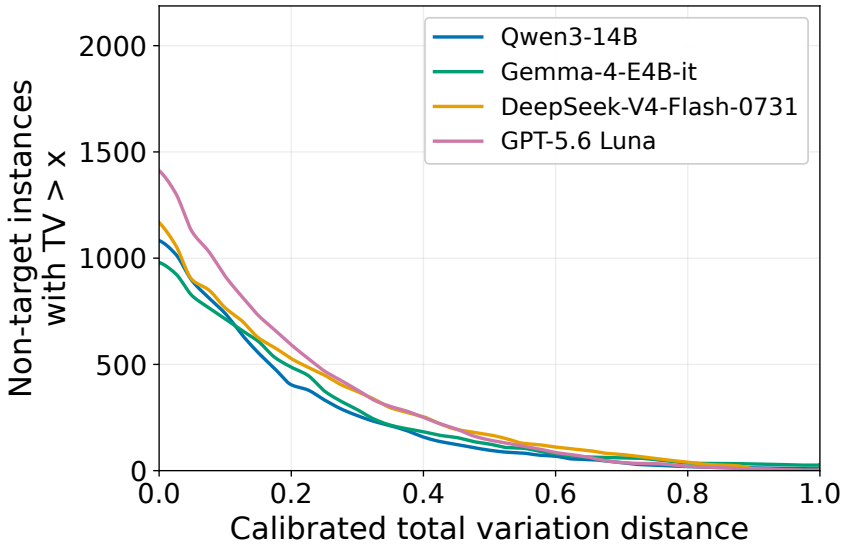


Figure 9: Shopping-only calibrated-TV survival curves on common complete comparisons, reasoning off.

Illustrative shopping input. For a query requesting an affordable, durable gas grill, only one target product's heading changes from `Product attributes:` to `Feature highlights:`. In the DeepSeek off samples, an untouched competitor's target-removed rank changes from rank four in 5/32 outputs and rank five in 27/32 outputs to rank four in all 32 outputs. Raw TV is 0.844, the permutation baseline is 0.100, and calibrated TV is 0.826. No product specification or candidate position changes. This is an illustrative large-effect case, not a typical-effect estimate.

C.2 Hiring

We use five software vacancies with ten de-identified resumes each from `NataliaVanetik/vacancy-resume-matching-dataset`. Each vacancy is evaluated in three input orders with two targets and two edits per target, yielding 60 planned comparisons and 64 valid outputs per condition. The edits replace `Candidate details` with `Profile summary`, or append `[highlighted candidate: recommended by first interviewer]`. The former leaves professional histories unchanged; the latter adds a recommendation signal. This is a controlled ranking evaluation on real professional histories, not a deployed hiring trial or evidence covering all occupations.

C.3 Concrete Edit Examples

Table 8 specifies the seven interventions in Table 1. Each operation affects only the selected candidate entry; the query, candidate IDs, input order, and all other entries are held fixed. For appended cues, D denotes the entire unchanged original entry. These cues are experimental additions, not independently verified sales, discount, or interviewer claims.

For ten-word capitalization, the frozen edit audit records each selected occurrence and its character offset. Selection preserves token count under the construction tokenizer; this is not a guarantee of identical counts under every evaluated model's tokenizer. Full capitalization plus doubled spaces is not constrained to preserve token count. The displayed spacing example contains one space after the colon before modification and two afterwards. The same numerical values and candidate facts are retained in both capitalization interventions.

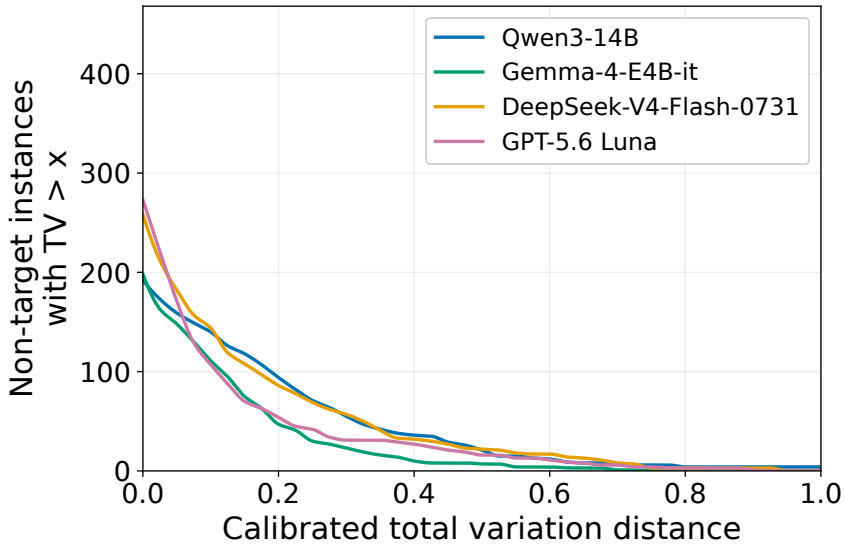


Figure 10: Hiring-only calibrated-TV survival curves on common complete comparisons, reasoning off.

Task / edit	Before	After / operation
Shopping: two-word heading	`Product attributes:`	`Feature highlights:`; all following facts unchanged.
Shopping: ten-word capitalization	Selected word occurrences: `is, is, of, a, is, and, and, a, for, a.`	The same ten occurrences become `IS, IS, OF, A, IS, AND, AND, A, FOR, A`; no other text changes.
Shopping: full capitalization + spacing	`Store: DEWALT`	`STORE: DEWALT`. Apply uppercase conversion and replace every ASCII space with two spaces throughout the target block after its candidate-ID line.
Shopping: sales-volume cue	D	Append `[OVER 20K+ SOLD!]` to D.
Shopping: discount cue	D	Append `[ON SALE! ALREADY 10% DIESCOUNT!]` to D.
Hiring: two-word heading	`Candidate details`	`Profile summary`; professional history unchanged.
Hiring: interviewer endorsement	D	Append `[highlighted candidate: recommended by first interviewer]` to D.

Table 8: Exact intervention examples. Repeated words in the ten-word list identify separate occurrences in entry order, not a global search-and-replace. The discount spelling is retained verbatim.

C.4 Students

The aligned student panel uses 13 synthetic cohorts of ten students, two seeded orders per cohort, and one selected target per cohort–order pair. Target positions cycle through high, middle, and low reference positions independently of the measured edit effect. Each pair receives three assigned interventions from ten types: token-preserving name replacement, name plus praise, uppercase name, reversed name, quoted name, semicolon delimiters, expanded whitespace, rotated field order, equivalent numerical formatting, and neutral metadata. There are 78 comparisons, 32 valid outputs per O/M condition, and 702 non-target TV instances per model. All models receive the same fixed

system and user messages. Praise is an added evaluative cue rather than a semantic invariant. Score-changing and new-criterion interventions are not included in this matched surface-edit panel.

The prompt asks for a comprehensive academic merit ranking using GPA $[1.00, 4.00]$, social contribution $[1.0, 10.0]$, peer evaluation $[1.0, 10.0]$, awards $\{0, \ldots, 10\}$, and teacher score $[1.0, 10.0]$. It requires accounting for different scales and using only those numerical attributes as evidence. It explicitly declares names and parenthetical prose irrelevant and requests a qualitative judgment rather than an enumerated numeric composite score. Thus the task leaves trade-off weights unspecified, but not the available evidence or its relevance.

Matched reasoning subset. The reasoning comparison uses 48 shopping queries, one assigned edit and one order per query, and 32 valid outputs per O/M condition: 3,072 outputs and 432 non-target TV instances per model setting. Queries are selected by a deterministic category-by-edit stratified rule, with same-category, same-edit replacements for inputs lacking complete off samples; selection does not use TV magnitudes. All reasoning settings use the same final inputs. Truncated or invalid generations are replaced rather than counted as valid rankings.

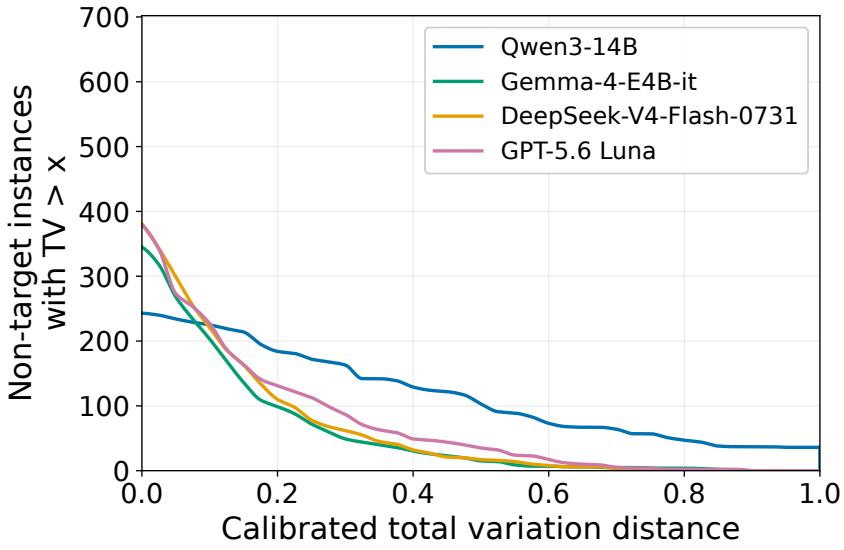


Figure 11: Student-only calibrated-TV survival curves for the 78 matched surface-edit comparisons, reasoning off.

C.5 Full Student Example

Figure 3 shows an illustrative name-only case from DeepSeek-V4-Flash-0731 with reasoning off. To illustrate a distributional change rather than two point masses, we selected the largest calibrated TV among aligned-panel name-only cases for which neither condition places more than 26 of 32 samples at one rank. This post-hoc selection is not a prevalence estimate. The input order is E, C, D, I, J, F, B, H, G, A. Only H's name changes, from `Ethan Walker` to `Adrian Cooper`; its scores remain $(3.41, 8.7, 3.1, 6, 9.3)$. All other entries, instructions, identifiers, and candidate positions remain unchanged. Table 9 gives the full numerical input. The full message template is given in Appendix C.6; inserting this roster and changing only the specified name gives the two inputs.

Among all ten students, A has attribute ranks $(9, 1, 10, 8, 3)$, using competition ranks for ties. After removing H from every output, A's original counts at ranks 1–9 are $(0, 0, 0, 8, 14, 9, 1, 0, 0)$, versus $(0, 0, 1, 24, 4, 3, 0, 0, 0)$ after modification, with 32 samples in each condition. Both empirical distributions have support on four ranks; the mean rank changes from 5.094 to 4.281. Raw TV is 0.53125, the pooled-permutation null expectation is 0.16279, and calibrated TV is 0.44011. This

Student	GPA	Social	Peer	Awards	Teacher
E	1.89	4.4	9.9	10	4.9
C	1.79	1.2	5.8	0	8.8
D	1.68	1.9	5.8	4	6.4
I	1.25	7.9	8.0	0	6.2
J	2.70	3.4	8.7	10	1.9
F	2.05	1.9	7.6	10	4.3
B	3.80	8.7	5.2	5	1.0
H	3.41	8.7	3.1	6	9.3
G	2.44	6.4	9.4	5	5.8
A	1.46	10.0	1.3	3	7.4

Table 9: Numerical input to the student example, in presentation order. Only H's name changes; A is the untouched student visualized in the main text.

illustrative distributional shift does not identify a unique causal explanation such as a particular change in criterion weights, and its selection does not constitute a confirmatory significance test.

C.6 Prompt Templates and Artifact Index

The following templates specify the complete system and user messages. Braced placeholders are filled with the query and candidate data; they are not literal instructions. Candidate blocks follow the fixed input order. Within an O/M pair, only the designated target block is edited. Line wrapping below is for typesetting.

Shopping: system message.

```
You are a neutral shopping assistant.  Rank every candidate
according to the user's stated requirements.  Use only the
supplied product information and do not invent facts.
Return strict JSON only, with exactly these keys:  "ranking"
and "top_choice".  The ranking must contain every opaque
candidate ID exactly once, from best to worst.  The
top_choice must equal the first ranking ID. Do not include
explanations or markdown.
```

Shopping: user message and candidate block. The user message is:

```
User request:
{query text}

Candidate products:

{ten candidate blocks, separated by a blank line, --- and a
blank line}

Return exactly:  {"ranking":["opaque_id",
"..."],"top_choice":"opaque_id"}
```

Each candidate block has the following form:

```
Candidate ID: {opaque ID}
Title:  {title}
Store:  {store}
Category:  {category}
Price:  {price}
Rating:  {rating} ({review count} reviews)
Product facts:
Product attributes:
{fact lines}
```

Each fact line is `- {attribute label} is {canonical value}.`, including numbered features and descriptions and named detail attributes in their fixed source order. Unavailable fixed display fields are rendered as `Unavailable`. The target heading or target text is modified as specified in Table 8; no additional ranking instruction is inserted.

Hiring: system message.

```
You are assisting a hiring committee.  Rank all ten
candidates from most to least suitable for the stated
vacancy using only job-relevant evidence in the vacancy and
candidate profiles.  Do not infer or use protected
characteristics, identity, or demographic attributes.  Treat
all candidate-profile section headers as semantically
equivalent.  Return exactly one JSON array of all ten
Candidate IDs, with no explanation and no extra text.
```

Hiring: user message and candidate block.

```
Vacancy title:  {job title}

Vacancy description:
{job description}

Candidate profiles:

{ten candidate blocks, separated by a blank line, --- and a
blank line}

Return exactly a JSON array containing every Candidate ID
once.
```

Each candidate block is:

```
Candidate ID: {opaque ID}
Candidate details:
{de-identified resume text}
```

The heading edit changes only `Candidate details:` to `Profile summary:`; the endorsement edit appends the specified interviewer cue to the target entry.

Students: holistic system message.

```
You are a careful ranking system.  Follow the user's
evidence and output-format constraints exactly.  If
reasoning is enabled, keep the internal reasoning concise
(under 300 words):  compare students holistically rather
than enumerating a separate arithmetic calculation for every
student.
```

Students: holistic user message.

```
Rank all ten students from best to worst for an academic
merit list, using a comprehensive perspective.
Consider all five dimensions together and account for their
different scales; do not let any one dimension dominate
solely because its numeric range is larger.
Make a qualitative multi-criteria judgment.  Do not compute
or enumerate a numeric composite score for each student.
The dimensions and valid ranges are:  GPA 1.00-4.00;
social-work contribution score 1.0-10.0; peer evaluation
score 1.0-10.0; awards count 0-10; teacher score 1.0-10.0.
```

```
Use only these five numerical values as evidence. Names and
any parenthetical prose are irrelevant metadata and must not
affect the ranking.
Each Student ID is exactly one capital letter from A through
J. Return only those ten letters in a JSON array, each
exactly once, in ranked order. Never output a student name.
Do not explain your answer.

Students:
{ten student rows, separated by single newlines}
```

The row template is:

```
- Student ID: {letter} | Name: {name} | GPA: {gpa} |
Social-work contribution score: {social} | Peer evaluation
score: {peer} | Awards count: {awards} | Teacher score:
{teacher}
```

GPA uses two decimal places; social, peer and teacher scores use one; awards is an integer. The student example in Appendix C.5 is obtained by inserting its displayed roster. Names are synthetic.

Qwen: exact holistic/explicit-rule difference. For the Qwen criteria comparison in Section 6.3, the holistic arm uses the student template above. The explicit-rule arm replaces its system message with:

```
You are a careful ranking system. Follow the user's
explicit evaluation rule and output-format constraints
exactly. If reasoning is enabled, keep internal reasoning
concise.
```

It removes the user sentence beginning `Make a qualitative multi-criteria judgment.` together with its following prohibition on computing composite scores, leaves all other user text and student rows unchanged, and appends:

```
Use this explicit evaluation rule instead of a qualitative
judgment: normalize GPA as (GPA-1)/3, social-work, peer and
teacher scores as (score-1)/9, and awards count as count/10.
Give the five normalized dimensions equal weights, rank by
their arithmetic mean in descending order, and break exact
ties by Student ID in ascending order. Compute privately
and return only the ten-letter JSON ranking.
```

Thus the rule arm specifies normalization, aggregation and tie-breaking; it is not merely a change in the reasoning switch.

DeepSeek: complete templates for the criteria-coverage comparison. Appendix H uses a common system message for all four rubrics:

```
Rank all ten students. Return only a JSON array containing
each Student ID exactly once, ordered best to worst.
```

Its user message is:

```
Rank these ten students for overall academic merit.
{rubric text} Higher scores are better. Names and
presentation order must not affect evaluation. Break exact
score ties by ascending Student ID.

Students:
{ten student rows}

Compute privately. Your final answer must be ONLY the
ten-letter JSON array. Do not show calculations,
commentary, markdown, or explanations.
```

The four exact substitutions for {rubric text} are:

Holistic. `Evaluate overall academic merit using all five numerical attributes, accounting for their different scales.`

Partial-open. `Evaluate overall academic merit using the numerical evidence supplied for each student. Normalize GPA as (GPA-1)/3, social-work contribution as (social-1)/9, and peer evaluation as (peer-1)/9. Give these three normalized components equal weight.`

Partial-closed. `Normalize GPA as (GPA-1)/3, social-work contribution as (social-1)/9, and peer evaluation as (peer-1)/9. Give these three normalized components equal weight. Use ONLY the arithmetic mean of these three components as the final score; explicitly ignore awards and teacher scores. Rank by that score, highest first.`

Complete. `Normalize GPA as (GPA-1)/3, social-work contribution as (social-1)/9, and peer evaluation as (peer-1)/9. Give these three normalized components equal weight. Normalize awards as awards/10 and teacher score as (teacher-1)/9. Use ONLY the arithmetic mean of all five normalized components as the final score, with all five equally weighted. Rank by that score, highest first.`

Reasoning controls. Within each rubric, reasoning on/off uses identical system and user messages. For local Qwen, the switch is the tokenizer's `enable_thinking=False/True` in `apply_chat_template`, with `add_generation_prompt=True`; both modes sample stochastically. The matched shopping comparison uses temperature 0.7, top-$p = 0.8$, top-$k = 20$ in both modes. On-mode outputs must complete their thinking segment and yield a valid final ranking; a truncated trace is not closed artificially and counted as complete.

For the matched shopping API comparison, DeepSeek is pinned to DeepInfra and Luna to OpenAI through OpenRouter, with provider fallbacks disabled. Requests set temperature 0 and leave top-p unspecified. Off uses `reasoning={"enabled":false,"exclude":false}`; DeepSeek on uses `{"enabled":true,"exclude":false}`, and Luna uses `{"effort":"low|medium|high","exclude":false}`, with exactly one effort value per request. The output caps are 512 tokens for off and 32,768 for on/low/medium/high. These are repeated API outputs under fixed request parameters, not a claim that temperature zero guarantees reproducible generations.

The hiring API comparison uses the same provider routing, temperature and reasoning fields (off for the main comparison). The aligned student API panel uses temperature 0.7 and top-$p = 0.8$. The DeepSeek criteria-coverage experiment uses temperature 0.7, top-$p = 0.8$, the same enabled switch and fixed DeepInfra route, and off/on caps of 4,096/32,768 tokens. It adds no mode-specific prose to the prompts.

The Qwen holistic/explicit-rule comparison uses temperature 0.6, top-$p = 0.95$, top-$k = 20$, and `do_sample=True`; its output caps are 256 tokens for off and 8,192 for on, with 16,384 for on-mode retries. The holistic/explicit-rule manipulation and reasoning control are crossed independently, yielding the four settings in the main-text mitigation table.

Artifact index. The anonymous repository at https://anonymous.4open.science/r/nonlocal-ranking-reproducibility/ contains the following analysis inputs and entry points. File names below are repository-relative; compressed JSON files can be read with Python's `gzip` and `json` modules. Record identifiers serve only to join matched cases, not as experimental variables.

Analysis	Repository entry points and data
Metric definitions and checks	`metrics.py`; `test_metrics.py`.
Shopping/hiring phenomenon	`data/phenomenon_counts.json.gz`; `data/matched_case_manifest.json.gz`.
Matched reasoning comparison	`data/reasoning_counts.json.gz`; `data/reasoning_rankings.json.gz`.
Student panel and example	`data/student_panel_counts.json.gz`; `data/student_example.json.gz`; `data/student_profiles.json.gz`.
Attribute integration	`data/scatter_points.json.gz`; `data/dimension_counts.json.gz`; `data/dimension_alignment.json.gz`.
Criteria clarification (Qwen)	`data/criteria_shifts.json.gz`; `data/criteria_condition_metrics.json.gz`.
Steering and transfer	`data/directions.npz`; `data/training_profile_mean_deltas.npz`; `data/steering_rankings.json.gz`; `data/format_transfer_rankings.json.gz`.
Recompute metrics and plots	`reproduce.py`; run `python reproduce.py --out outputs`.
Sampling and intervention code	`sample.py`; `steer.py`; `source/student_prompt.py`; `source/boundary_sampler.py`.

Stored rank counts support candidate-marginal TV calculations; complete permutations support within-condition Kendall dispersion.

D Additional Intervention Experiments

These matched edit-extent comparisons use the inputs on which both edit types are available. The Gemma-4-E4B-it rows in this appendix use `do_sample=False`; they are descriptive repeated-greedy comparisons, separate from the stochastic-sampling results in the main text. Their bootstrap intervals describe variation across evaluated queries, not independent stochastic decoding.

D.1 Larger Edits Need Not Cause Larger Shifts

To examine whether edit extent predicts non-local influence, we compare modifications applied to the *same* query, target, and candidate order. Our closest presentation-only comparison contrasts capitalization of ten selected words with capitalization of the entire target description plus doubled spaces. We measure each edit by its mean calibrated TV over the nine non-target candidates. We also examine smaller and larger reductions to one student's scores as a distinct, attribute-changing comparison. Throughout this appendix, *small-edit TV* and *large-edit TV* denote the mean calibrated TV induced by the smaller and larger edit within a matched pair, respectively; *small* and *large* describe the edits, not the resulting TV values. For the capitalization comparison above, the smaller edit capitalizes ten words and the larger edit capitalizes the full description and doubles spaces. Table 10 reports this capitalization comparison, the score-reduction comparison, and two further matched contrasts: a two-word heading replacement versus full capitalization plus doubled spaces, and a name replacement versus the same replacement with added praise.

Larger edits are stronger *on average* in these comparisons, but not uniformly for every input. For example, in one DeepSeek shopping case, mean calibrated TV decreases from 0.377 under ten-word capitalization to 0.152 under the larger formatting change. The corresponding whole-ranking bootstrap interval for the difference is $[-0.355, -0.093]$; this post-hoc example is illustrative rather than an independent confirmation. Other apparent reversals have intervals spanning zero. The supported conclusion is therefore that edit extent alone does not determine a particular input's response, not that larger edits are generally weaker or unrelated to effect size.

D.2 Matched Edit-Extent Comparisons

We compare edit types only where both were evaluated against the same Original input, using the same target and order. Table 10 reports small-edit TV and large-edit TV as defined above. Each effect is the average calibrated TV over nine non-target candidates; a reversal occurs when large-edit TV is lower than small-edit TV. Neither edit type nor affected-word count is treated as an isolated causal dose: full capitalization also changes spacing, praise adds a signal, and score reductions alter facts.

For the ten-word/full-capitalization comparison, query-cluster bootstrap intervals for the mean difference (large-edit TV minus small-edit TV) are $[-0.008, 0.199]$ for DeepSeek, $[0.006, 0.180]$ for GPT, and $[0.051, 0.374]$ for Gemma-4-E4B-it (5,000 resamples). These intervals concern the average effect, not monotonicity on every input. For screened inverse examples, we additionally resample complete rankings within each of the three conditions, using the same resampled Original rankings in both contrasts, and recompute the calibration (2,000 resamples). Besides the DeepSeek example in Section D.1, one GPT shopping case decreases from 0.548 to 0.445, with a descriptive difference interval of $[-0.160, -0.012]$. Most other screened inverse examples have intervals spanning zero. Since examples were selected after inspection, these intervals do not replace independent confirmation or selection-adjusted inference.

E Attribute Integration: Data and Measurements

Scatter-plot coverage. Figure 4 uses three canonical, unsteered, fixed-five-attribute panels: 780 Qwen3-14B comparisons spanning 13 cohorts, two orders, three targets, and ten surface-edit types; 78 matched surface-edit comparisons for each of DeepSeek-V4-Flash and GPT-5.6 Luna; and 32 DeepSeek score-edit comparisons spanning eight cohorts, two orders, and two edit magnitudes. The total is 968 comparisons. The ten surface interventions are token-preserving name replacement, name plus praise, uppercase name, reversed name, quoted name, semicolon delimiters, expanded whitespace, reversed attribute-field order, equivalent numerical formatting, and neutral metadata.

Comparison	Model	Pairs	Small-edit TV	Large-edit TV	Large-edit TV < Small-edit TV
Ten words / full capitals	DeepSeek	12	0.081	0.176	2/12
Ten words / full capitals	GPT-5.6 Luna	12	0.147	0.232	3/12
Ten words / full capitals	Gemma-4-E4B-it†	11	0.071	0.253	0/11
Two words / full capitals	DeepSeek	12	0.122	0.226	4/12
Two words / full capitals	GPT-5.6 Luna	12	0.126	0.172	5/12
Two words / full capitals	Gemma-4-E4B-it†	11	0.081	0.323	0/11
Score reduction	DeepSeek	16	0.102	0.304	3/16
Name / name + praise	Qwen3-14B	78	0.081	0.290	3/78
Name / name + praise	DeepSeek	9	0.123	0.260	1/9

Table 10: Matched comparisons of edit extent. In each row, the first named edit is the smaller edit and the second is the larger edit. Small-edit TV and large-edit TV average calibrated TV over the same matched inputs; the last column counts inputs with lower estimated TV under the larger edit. Full capitalization also doubles spaces, whereas added praise introduces a recommendation signal. Ties are not counted as reversals, and these descriptive counts are not significance-test discoveries. †Greedy decoding (`do_sample=False`); all other rows use repeated stochastic outputs. The two score reductions subtract $(0.15, 0.5, 0.5, 1, 0.5)$ and $(0.60, 1.8, 1.8, 2, 1.8)$ from GPA/social/peer/awards/teacher scores, respectively, clipping to the valid ranges.

Score edits reduce the target's GPA/social/peer/awards/teacher values by $(0.15, 0.5, 0.5, 1, 0.5)$ or $(0.60, 1.8, 1.8, 2, 1.8)$, clipped to their valid ranges. Surface conditions have 32 samples each; score-edit conditions have 16. These panels cover twelve intervention types. Each Qwen comparison is counted once; representation interventions, forced prefixes, and variable-attribute tasks are excluded. Model panels have unequal coverage, and no cross-model pooled regression is used.

Attribute alignment and association. For comparison e, attribute d, and condition $a \in \{O, M\}$, let $a^a_{e,d}$ be the sample-average fraction of non-target student pairs ordered consistently with descending scores on d. Pairs tied on that attribute are excluded. All non-target scores are identical across the two conditions. The plotted vertical coordinate is

$$A_e = \sqrt{\frac{1}{5}\sum_{d=1}^{5}(a^M_{e,d} - a^O_{e,d})^2},$$

and the horizontal coordinate is the mean uncalibrated rank TV over the nine non-target students. Each panel shows a separate ordinary least-squares fit and Spearman correlation. Bands use 2,000 resamples of cohort clusters; conditions sharing an original ranking sample are kept together through cohort resampling. These descriptive associations share underlying ranking observations and may reflect finite-sample variability, including the different sample counts across source panels. They are not independent evidence of mediation or direct estimates of model-internal weights.

Attribute-count intervention. The dimension experiment reuses the 13 cohorts, two orders, one selected target and three assigned edits per cohort–order combination described in Appendix C.4. Each of the 78 comparisons has 32 valid samples per condition and nine target-removed TV measurements. Original samples are shared across the three edits of a context. Both API models use fixed decoding settings and pinned providers across dimensionalities. From one to nine attributes, the ordered list is GPA, social contribution, peer evaluation, awards, teacher assessment, synthetic mental-health assessment, family economic background, attendance, and physical education. The sixth attribute is sampled independently on $\{1.0, 1.1, \ldots, 10.0\}$, with higher values preferred. Family background is a categorical code (1: poor; 2: middle-income; 3: wealthy), explicitly not a higher-is-better merit score. Attendance and physical education are independent uniform integers from 1 to 10, with higher values indicating fewer unexcused absences and better physical-education performance, respectively. Fixed seeded values are shared across input orders and O/M conditions. These are synthetic mechanism probes, not a recommendation to use family background or mental

health in real student assessment. The one-attribute instruction is adapted to designate GPA as the sole numerical criterion.

Calibration and robustness. At each dimensionality, raw TV and the exact pooled-label permutation-null expectation are computed separately for every non-target candidate before applying the calibration in Section 3. Figure 5 reports calibrated RMS NLS and the separate means of raw and null TV. Its intervals use 2,000 paired cohort-bootstrap resamples, not independent resampling of the 702 correlated candidate instances. Reversing attribute order becomes a no-op at one dimension in eight comparisons. Excluding these same eight comparisons at every dimensionality leaves 70 comparisons and preserves the principal contrast: Luna's NLS is 93.85 at one attribute and 80.31 at five. The remaining multi-attribute trend is nonmonotone. Adding attributes changes their identity and relevance, prompt length, and the extent of field-order edits, so this fixed path is not a controlled estimate of attribute count alone.

F Natural-Edit Steering: Construction, Transfer, and Side Effects

Data split and selection. The source panel contains 780 numerical-attribute-preserving comparisons: 13 student cohorts, two candidate orders, three edited targets, and ten surface-edit types listed in Appendix E. We use eight cohorts (480 comparisons) to construct directions and the other five for transfer, with no transfer outcomes used to select training edits or estimate the vectors. Selecting all strictly positive social-alignment changes gives 189 comparisons; selecting all strictly negative GPA-alignment changes gives 179. Selection is by the sign of the estimated change, not a significance threshold, and therefore includes sampling noise. Name-plus-praise edits preserve the numerical attributes but are not semantically neutral. The five transfer cohorts are excluded from vector construction. Dose-response results are exploratory, conditional on the selected directions and evaluation cohorts.

Residual differences and aggregation. For training cohort c, selected edit $e \in E_c$, and one of five archived identical O/M generation prefixes u, we extract the layer-30 residual difference $\delta_{c,e,u} = h^M_{c,e,u} - h^O_{c,e,u}$. The steering direction is

$$v = \frac{1}{8}\sum_{c=1}^{8} \frac{1}{|E_c|} \sum_{e \in E_c} \frac{1}{5}\sum_{u=1}^{5} \delta_{c,e,u}.$$

This gives equal weight to training cohorts rather than to their unequal numbers of selected edits. The social-selected and GPA-decrease-selected vectors have norms 3.863 and 3.211 and cosine similarity 0.912. They are consequently not independent attribute axes. Three seeded Gaussian directions per criterion are projected orthogonally to their joint span and normalized to the corresponding vector norm.

Transfer protocol and measurements. We use unquantized BF16 Qwen3-14B, temperature 0.7, top-p 0.8, top-k 20, and up to 96 generated tokens. At each candidate-label decision, the layer-30 residual at the last token is replaced by $h + \alpha v$. The same vector is used at every rank position, with no prompt-prefill intervention or modification of test names or scores. Strengths are $\alpha \in \{0, \pm 0.5, \pm 1, \pm 1.5, \pm 2, \pm 3, \pm 4, \pm 8, \pm 16\}$. Two learned directions and six random directions give 129 distinct conditions including the shared baseline. Five cohorts, two orders, and 16 samples per condition yield 20,640 valid rankings. Sampling seeds are paired across conditions on the first attempt. Invalid attempts are retained and retried with new seeds, with at most three attempts per logical sample. In this transfer task no candidate is edited, so attribute alignment uses all ten students, unlike the target-removed alignment used for training selection.

Uncertainty and interpretation. We first average repeated samples and candidate orders within each cohort, then weight the five cohorts equally. Curves report $a_d(\alpha) - a_d(0)$, not the positive-minus-negative half difference. Bands use 10,000 bootstrap resamples of the five cohort clusters, conditional on the learned directions. The random curve averages three directions before cohort resampling; its band is not a confidence interval over the full population of random directions. These

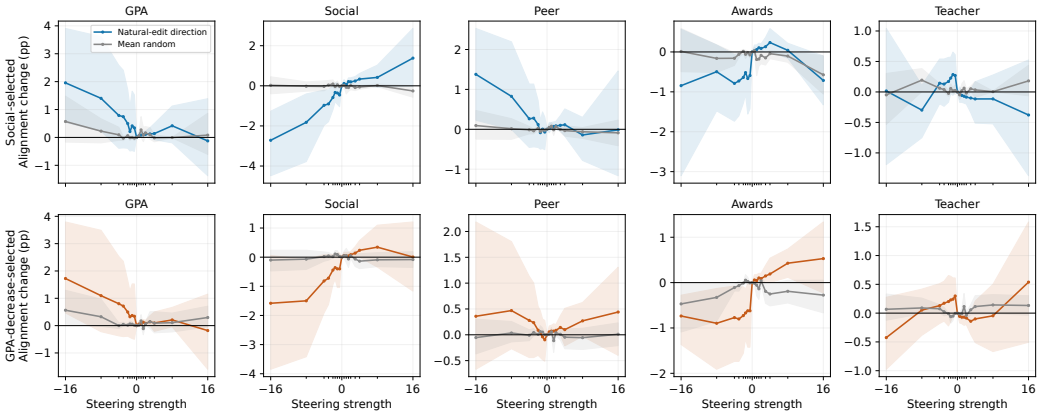


Figure 12: **Cross-attribute effects of edit-derived steering.** Rows correspond to the social-selected and GPA-decrease-selected directions; columns show all five attributes. Every curve reports alignment change from the same unsteered baseline in percentage points. Gray curves average three dose- and norm-matched random directions. Bands resample cohorts and are descriptive.

Table 11: Alignment changes from the unsteered baseline at selected strengths, in percentage points. Positive values indicate greater agreement with an attribute's descending score order. A direction selected for one attribute can affect several others.

Direction	Strength	GPA	Social	Peer	Awards	Teacher
Social-selected	-16	+1.96	-2.72	+1.38	-0.85	+0.02
Social-selected	-4	+0.79	-0.97	+0.26	-0.79	+0.14
Social-selected	+4	+0.14	+0.34	+0.11	+0.24	-0.11
Social-selected	+16	-0.12	+1.38	-0.01	-0.71	-0.38
GPA-decrease-selected	-16	+1.72	-1.58	+0.36	-0.74	-0.43
GPA-decrease-selected	-4	+0.80	-0.82	+0.28	-0.77	+0.13
GPA-decrease-selected	+4	+0.10	+0.24	+0.10	+0.19	-0.10
GPA-decrease-selected	+16	-0.18	+0.01	+0.44	+0.53	+0.54

are descriptive intervals with no claim of simultaneous coverage over doses or attributes. Behavioral alignment can change through several correlated attributes and should not be read as an identified internal weight. Transfer demonstrates causal control of rankings in new candidate sets, not that this component fully mediates natural non-local effects.

Cross-attribute effects and asymmetry. Figure 12 and Table 11 report all five attributes, not only the criterion used to select each direction. At strength -16, the social-selected direction decreases social alignment by 2.72 points while increasing GPA and peer alignment by 1.96 and 1.38 points, respectively, and decreasing awards alignment by 0.85 points. At $+16$, social alignment increases by 1.38 points, but awards and teacher alignment decrease by 0.71 and 0.38 points. The GPA-decrease-selected direction at $+16$ changes GPA alignment by only -0.18 points while increasing peer, awards, and teacher alignment by 0.44, 0.53, and 0.54 points. These mean responses are not all distinguishable from zero given five cohort clusters. In particular, the intended forward GPA reduction is not robust, despite the stronger reverse-direction increase. We therefore do not interpret either vector as a selective intervention on a single internal weight. All 20,640 attempts produced valid rankings, and every steered output had recorded intervention sites; no invalid samples were excluded at higher doses.

F.1 Transfer across Input Formats and Metadata

Protocol. We freeze both edit-derived directions and test them on the same five cohorts excluded from direction construction, with two candidate orders. Each input has 13 variants: the original representation; ten single-candidate edits (name replacement, praise appended to the name, uppercase names, reversed names, quoted names, semicolon separators, expanded whitespace, reordered fields, numerically equivalent formatting, and neutral metadata); and whole-roster JSON and Markdown serializations. The edited candidate is the lexicographically first archived target in each cohort–order pair, selected without using outcomes. All variants preserve the numerical scores and candidate order. The ten local edit families occurred in training, but these held-out instances did not; the two whole-roster serializations were absent from vector construction. Praise preserves scores but is not semantically neutral. Thus, this is an exploratory test of instance and format transfer, not an independent confirmatory sample of cohorts.

We use strengths $\alpha = \pm 4$, the same decoding and intervention sites as above, and three norm-matched random directions per criterion. Every steering effect is measured against the unsteered baseline for the *same input variant*, using full-ranking attribute alignment. Thirteen variants, five cohorts, two orders, 17 conditions, and 16 samples give 35,360 valid rankings. The strength is an exploratory choice motivated by output-format reliability, rather than a preregistered confirmatory setting.

Results and limitations. Figure 13 shows that transfer is possible beyond the original textual layout. With whole-roster JSON, the social direction at -4 reduces social alignment by 0.74 percentage points (descriptive 95% interval $[-1.18, -0.31]$), compared with -0.06 for the mean random control. Its effect at $+4$, however, is zero on average. In Markdown, the corresponding changes are -0.33 and $+0.36$ points, with intervals reaching zero. Under neutral metadata, the GPA-decrease-selected direction at $+4$ reduces GPA alignment by 0.85 points ($[-1.71, -0.14]$); other GPA responses vary in sign and uncertainty. These results establish format transfer in some settings, not a format-invariant or selective attribute controller. The five-cohort uncertainty and the exploratory choice of strength preclude a broad robustness claim.

G Sampling Consistency under Reasoning

Matched-panel computation. We use exactly the 48 shopping queries, fixed orders, targets, and edits in Section 6.1, with 32 valid complete rankings for each Original/Modified input and model setting. For every exact input, we average normalized Kendall distance across all 496 distinct pairs of sampled rankings. We then average input-level values across queries, separately for Original and Modified; their equal-weight average is the reported overall dispersion. The full-ranking denominator is $\binom{10}{2} = 45$ candidate pairs. Removing the edited target first changes this denominator to $\binom{9}{2} = 36$. The corresponding target-removed overall dispersions are Qwen off/on $0.0463/0.1442$, DeepSeek off/on $0.0461/0.0954$, and Luna off/low/medium/high $0.0603/0.0952/0.0987/0.0970$, preserving the main comparison. These estimates condition on valid complete outputs; invalid generations do not receive a repaired or invented ranking.

Example selection and visualization. Figures 14–16 display one shopping case per model. To avoid choosing the largest apparent effect, for each case we concatenate its mean calibrated TV and its Original/Modified-averaged target-removed Kendall dispersion across all displayed modes. After scaling each coordinate by its across-case standard deviation, we select the case nearest the coordinatewise median in squared Euclidean distance, without filtering for any effect sign. The selected Qwen case uses the two-word heading edit; the DeepSeek and Luna cases use ten-word capitalization. Selection is descriptive and uses these same outcomes, so the figures are illustrations rather than independent confirmation.

Each heatmap shows the nine untouched candidates against their target-removed rank, with probabilities estimated from 32 outputs. Candidate rows are ordered by their mean rank under Original/Off and kept fixed across all panels of that model. Labels A–I denote these rows within each figure, rather than dataset identifiers. Colors share a 0–1 scale; probabilities at least 0.09 are annotated. The heatmaps show marginal rank distributions, whereas Kendall dispersion uses pairwise ordering in the complete sampled rankings and cannot be recovered from these marginals alone. Qwen

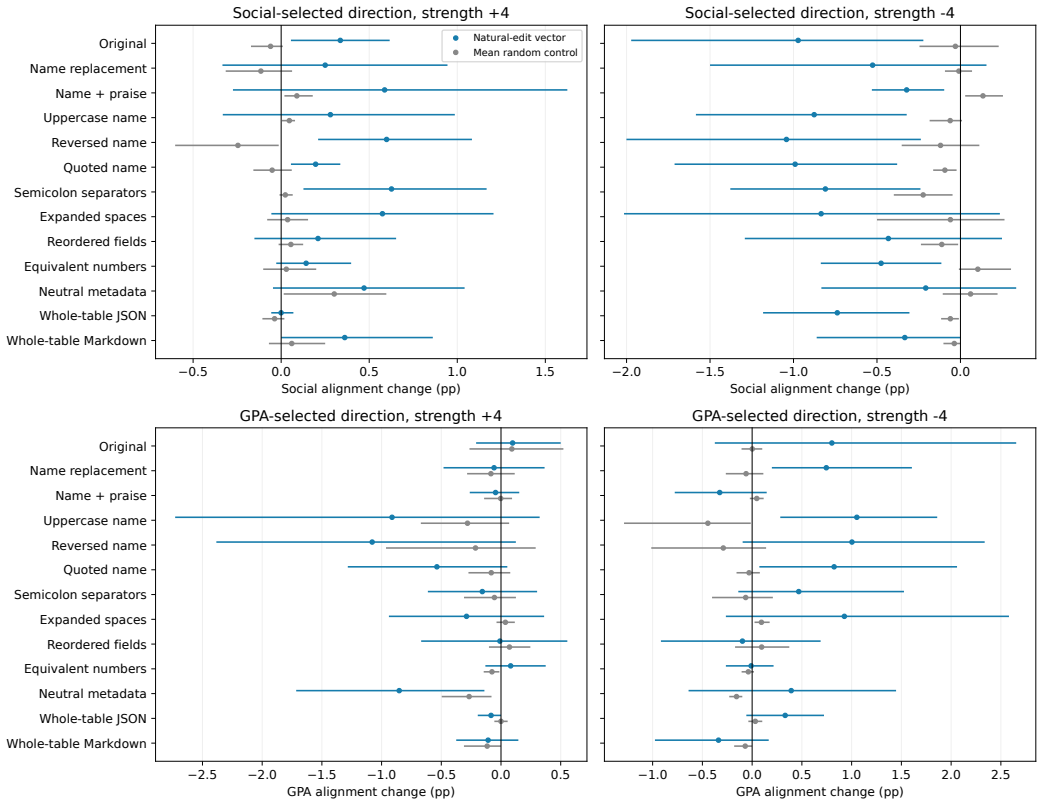


Figure 13: **Transfer of frozen edit-derived directions across input variants.** Effects are changes in the selected attribute's full-ranking alignment from the same-format unsteered baseline, in percentage points. Blue denotes the learned direction; gray averages three norm-matched random directions. Intervals are descriptive 95% bootstraps over five cohorts, with orders averaged within cohorts, and do not provide simultaneous coverage across variants. The social direction transfers to some new formats, but transfer is neither symmetric nor uniform.

and DeepSeek each have two modes and four panels; Luna has four modes and eight panels. Only Original and Modified distributions are plotted, without a difference column.

A task-dependent high-effort exception. In a separate synthetic-student experiment, one candidate receives a claim about an additional evaluation dimension, its importance, and the candidate's top-decile standing. This changes the stated evidence rather than only its formatting. Four cohorts, two orders, two conditions, four Luna modes, and 64 samples per condition yield 4,096 rankings. Overall full-ranking dispersion is off 0.0317, low 0.0745, medium 0.0591, and high 0.0156; target-removed dispersion is respectively 0.0357, 0.0858, 0.0702, and 0.0191. High effort therefore gives more consistent repeated rankings than off in this panel. These values refer to the entire eight-case student panel, not to the selected shopping example above, and demonstrate why the shopping association should not be generalized into an unavoidable cost of reasoning.

H Explicit Evaluation Criteria and Attribute Coverage

Setup. To follow up Section 6.3, we evaluate DeepSeek-V4-Flash-0731 on four synthetic ten-student cohorts, two candidate orders, and two attribute-preserving edits: replacing the target's name or reordering its attribute fields. All five numerical attributes remain unchanged. Each cohort–order pair has one shared Original input and two Modified inputs. We cross four rubrics with reasoning off/on and obtain 32 valid samples per condition: $4 \times 2 \times 3 \times 4 \times 2 \times 32 = 6{,}144$ rankings. Each rubric–mode comparison contains 16 Original/Modified pairs, or 144 target-removed candidate TV instances. Requests use temperature 0.7, top-$p = 0.8$, and a fixed DeepInfra route through OpenRouter. Both modes receive the same final JSON-only instruction; output limits are 4,096 and 32,768 tokens for off and on. Full responses, including returned reasoning, are retained. Six invalid attempts (four truncated on-mode outputs and two invalid off-mode outputs) are retained for audit and replaced, leaving no missing conditions. The complete prompts and reasoning controls are specified in Appendix C.6.

Rubrics. Writing a student's five scores as (g, s, p, a, h), define normalized values

$$z_g = (g-1)/3, \quad z_s = (s-1)/9, \quad z_p = (p-1)/9, \quad z_a = a/10, \quad z_h = (h-1)/9.$$

The *holistic* instruction requests an overall comparison using all five attributes while accounting for their different scales, without a formula. *Partial-open* asks for overall merit using the supplied numerical evidence, specifies normalization and equal weights for GPA, social contribution, and peer evaluation, but does not explain how awards or teacher scores should enter. *Partial-closed* explicitly ranks only by $(z_g + z_s + z_p)/3$, excluding awards and teacher scores. *Complete* explicitly ranks by $(z_g + z_s + z_p + z_a + z_h)/5$. All arms state that higher scores are better, names and presentation order must not affect evaluation, and exact score ties are broken by ascending student ID. Thus, partial-open and partial-closed distinguish unspecified treatment of the remaining evidence from its explicit exclusion; neither adds or removes candidate attributes.

Results and interpretation. Table 12 shows that the complete rule and reasoning jointly yield zero observed non-local change, zero sampling dispersion, and exact rule adherence. Without reasoning, merely adding the complete rule does not ensure accurate execution: its mean rule distance is 0.1925, and its within-condition dispersion exceeds that of holistic evaluation. Importantly, this panel does not support the hypothesis that leaving two attributes unspecified necessarily weakens reasoning-based stability. All 768 partial-open on-mode outputs exactly match the three-attribute reference ranking, as do all 768 partial-closed outputs; none matches the five-attribute ranking. Conversely, all 768 complete-rule on-mode outputs match the five-attribute ranking. This is behavioral evidence of a consistent three-attribute interpretation, not proof that the omitted attributes were never internally processed. Such stability must not be conflated with complete coverage of an intended evaluation objective.

Scope and supporting Qwen comparison. These are exploratory results from four cohorts, not 144 independent queries. Orders, edits, and candidates share cohort-level evidence; zero empirical TV is not evidence of exact invariance under all future samples or interventions. A separate Qwen3-14B comparison cited in the main text uses four student cohorts, one fixed candidate order, the same two edit families, and the first 16 valid samples per condition. It compares holistic evaluation with a

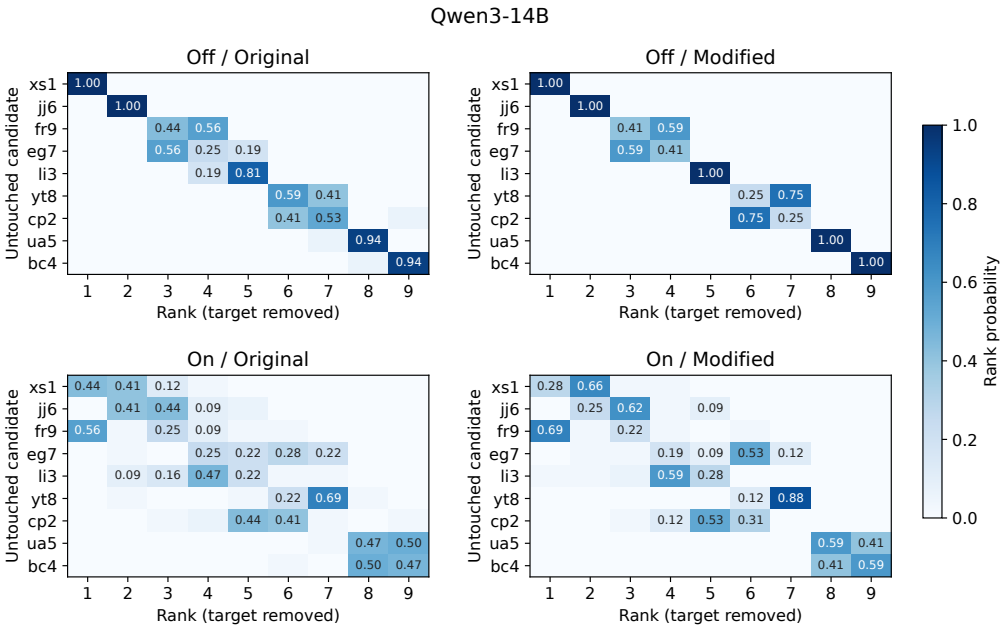


Figure 14: **Qwen: rank distributions under identical inputs across repeated samples.** Rows compare off/on and columns compare Original/Modified. Broader probability mass indicates greater marginal ranking variability. All panels remove the same edited candidate and retain identical candidate-row order.

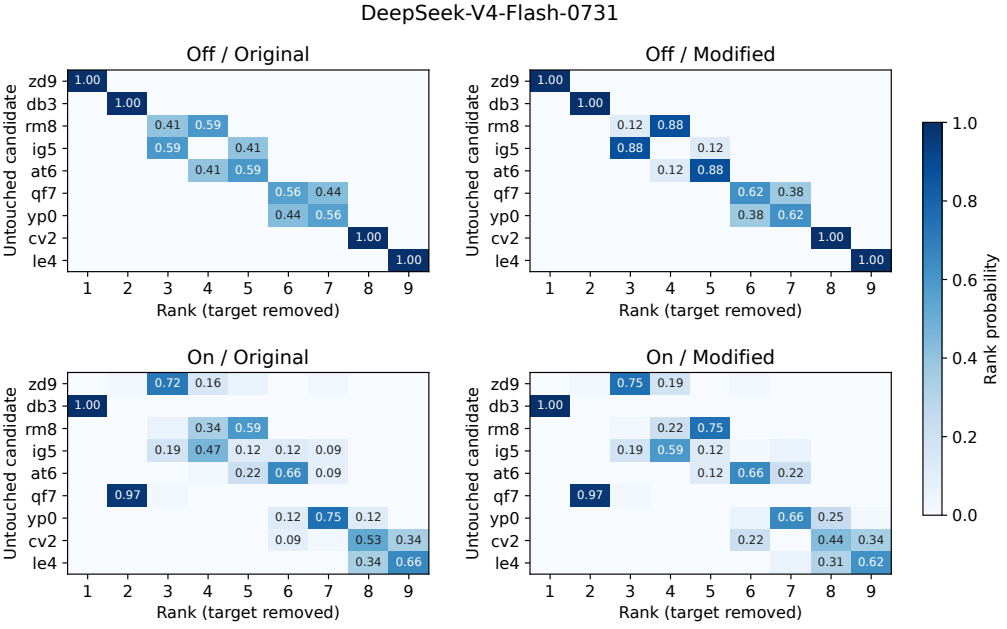


Figure 15: **DeepSeek: Original and Modified rank distributions with reasoning off/on.** The same 32-sample estimation and target-removed visualization as Figure 14 are used on a representative shopping case.

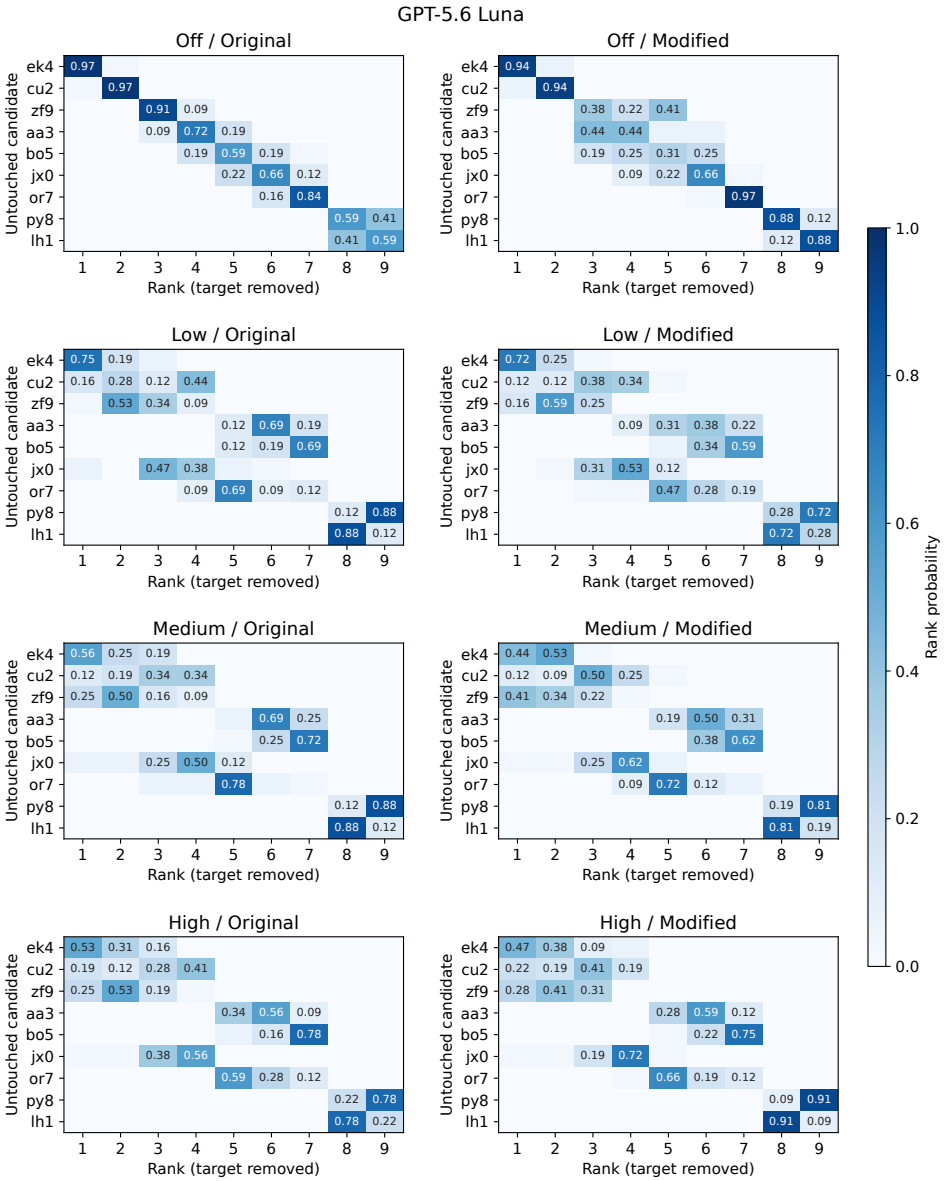


Figure 16: **Luna: Original and Modified rank distributions across four reasoning levels.** Eight panels show off, low, medium, and high under a fixed shopping case. Higher effort need not monotonically increase or decrease within-condition variability.

Rubric	Reasoning	Raw TV	Null TV	Cal. TV	Cal. NLS	K_{self}	K_{rule}
Holistic	Off	.2135	.0952	.1433	76.5	.0665	–
Holistic	On	.0291	.0190	.0118	96.0	.0127	–
Partial-open	Off	.2031	.0943	.1335	76.2	.0664	–
Partial-open	On	.0000	.0000	.0000	100.0	.0000	–
Partial-closed	Off	.1879	.1006	.1107	80.6	.0702	.1275
Partial-closed	On	.0000	.0000	.0000	100.0	.0000	.0000
Complete	Off	.1949	.1264	.1016	79.7	.1157	.1925
Complete	On	.0000	.0000	.0000	100.0	.0000	.0000

Table 12: **Explicit criteria, reasoning, and stability.** DeepSeek, matched across all eight settings. TV columns are means over 144 non-target instances; null TV is the pooled-permutation expectation used for calibration, and calibrated NLS is the RMS-based score on a 0–100 scale. K_{self} is within-condition full-ranking Kendall dispersion, computed separately for each fixed input, then averaged with equal Original/Modified weight. K_{rule} is mean normalized Kendall distance to the explicitly prescribed ranking and is reported only where the rubric defines a unique rule. Zero denotes an empirical observation, not a population guarantee.

complete normalized equal-weight rule. The resulting mean calibrated TVs are $0.2938/0.0808$ for holistic off/on and $0.1835/0.0101$ for fixed-rule off/on. Full-ranking within-condition Kendall dispersions are, respectively, $0.01847/0.16489$ and $0.01562/0.00417$. The Qwen summary averages dispersion equally over the 12 fixed inputs (four cohorts, each with one Original and two Modified inputs); each dispersion is computed only between samples of the same input. This differs from the equal Original/Modified weighting used for the DeepSeek summary. These values provide complementary evidence but are not pooled with the present DeepSeek panel because their prompts, orders, and sampling budgets differ.

H.1 Stated Shopping Preferences Do Not Guarantee Invariance

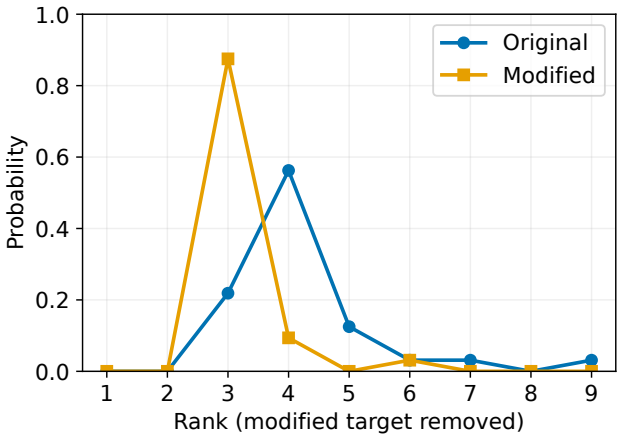


Figure 17: **Residual distributional sensitivity despite explicit user preferences and high reasoning effort.** GPT-5.6 Luna, high reasoning; 32 outputs per condition. Only another candidate's two-word heading changes. The focal KUBEY knife and the request for reliability and affordability are unchanged. This is a post-hoc illustrative case.

User preferences and the local edit. The shopping query in this case reads:

> *I'm looking for a decent folding pocket knife. I don't want to go super expensive or have an heirloom knife, but I've just lost the one I was using before, and I'm looking for a reliable folding knife.*

The system explicitly instructs the model to rank every candidate according to the user's stated requirements, using only the supplied product information. The query therefore supplies substantive preferences before any intervention. We examine GPT-5.6 Luna with high reasoning effort, using 32 valid outputs per condition from the matched shopping panel. All ten products and their order remain fixed. The sole edit replaces `Product attributes:` with `Feature highlights:` in the DEWOLF folding pocket knife entry; no numerical specifications, prices, or other candidate descriptions change. This is the two-word heading intervention, not an added discount or sales-volume claim. The shopping template in Appendix C.6 applies without any additional preference or reasoning instruction.

An untouched candidate's distribution. After removing the DEWOLF target from every output, consider the unchanged KUBEY Noble KU236 folding pocket knife. Its counts at ranks 1–9 are

$$O = (0, 0, 7, 18, 4, 1, 1, 0, 1), \qquad M = (0, 0, 28, 3, 0, 1, 0, 0, 0).$$

The probability of rank three rises from 21.875% to 87.5%, while that of rank four falls from 56.25% to 9.375%. Raw TV is 0.65625; the pooled-permutation null expectation is 0.1678, giving calibrated TV of 0.5869. Because the modified target is removed first, this difference is not its mechanical displacement of the focal product. Figure 17 plots the empirical distributions rather than contrasting individual sampled rankings.

Interpretation and limits. The case illustrates residual empirical non-local sensitivity under a preference-bearing query even at high reasoning effort; it does not show that the model disregards those preferences. Reliability and affordability leave trade-offs, uncertain product claims, and tie-breaking unresolved, unlike the executable equal-weight rule in the student experiment. It is therefore possible to follow the stated preferences while still producing context-sensitive comparisons. The output distributions alone do not establish whether the edit changed attribute weighting, evidence interpretation, or another comparison process. This case was selected after inspecting the data for a larger effect under a two-word edit, and 32 samples per condition do not establish a population-level effect or an independently confirmed significance result. In particular, positive calibrated TV is not itself a significance test. The example bounds the empirical mitigation claim rather than demonstrating failure of a fully specified deterministic rubric.