WHEN AGREEING CONTEXT HURTS: SEQUENCE PAIRING IN SUPERVISED FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidence in a neighboring contract can agree with a target judgment yet make a model less likely to answer it correctly. On ContractNLI, deleting that neighbor's target-consistent evidence corrects many pre-existing errors. We study how training-time sequence pairing shapes this behavior. Across supervised fine-tuning (SFT) arms, we keep complete judgments, answers, and optimizer-update memberships fixed while changing which judgments share a sequence. ProofWriter establishes the controlled comparison: pairing changes delivered answers, complete training layout changes the contrast, and 4B checkpoints selected by single-question accuracy trail alternatives by more than ten points on paired requests over a separate set of SFT-unseen theories. In ContractNLI, the original pairing rule deliberately guaranteed disagreement on the target hypothesis. Before new training, we predicted that weakening this relation would attenuate the observed context and deletion effects. Re-pairing the same judgments raises training agreement from zero to about 43%; in two independently trained repaired checkpoints, the roughly twenty-point accuracy deficits with agreeing versus conflicting neighbors move near zero, and deleting target-consistent rather than neutral text yields exactly zero change in target accuracy, while document-conditioned paired-task competence is retained. The repair also incurs condition-specific losses. Sequence partners can shape how fine-tuned models use neighboring information, and checkpoint comparisons should reflect the requests expected at use time.

1 INTRODUCTION

Each ContractNLI judgment asks about its own contract. Yet cross-contract-trained checkpoints answer a target question about twenty points less accurately when a neighboring contract agrees with the target answer than when it conflicts. Deleting target-consistent evidence from that neighbor, rather than matched neutral text, corrects many errors that were present before either edit. This pattern appears in two independently trained checkpoint pairs on ContractNLI (Koreeda & Manning, 2021) (Figure 1; a deletion case appears in Figure 5). Can the way complete judgments are paired during training teach a model how to use neighboring information later?

The same complete judgments, answers, and optimizer-update memberships appear in both ContractNLI training arms; what differs is which judgments share a sequence. We first establish this comparison in a controlled setting before returning to the natural-text finding and a training test of the relation imposed by its pairing rule.

Specifying complete judgments and their answers leaves another choice: which judgments become partners in an autoregressive sequence. An occurrence list does not determine that assignment. On ProofWriter (Tafjord et al., 2021), each occurrence is a complete world–query problem whose label in $\{\texttt{True}, \texttt{False}, \texttt{Unknown}\}$ is determined by that problem alone. We pair these occurrences within or across worlds, holding their answers, position roles, and update memberships fixed.

Partner assignment and row layout are separate choices. A shared answer block places both questions before both answers; interleaved turns supervise the first answer before the second question appears. The same partners can thus surround supervised answers differently. Section 2 formalizes these choices, and Figure 2 shows the controlled rewiring and the two layouts.

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.

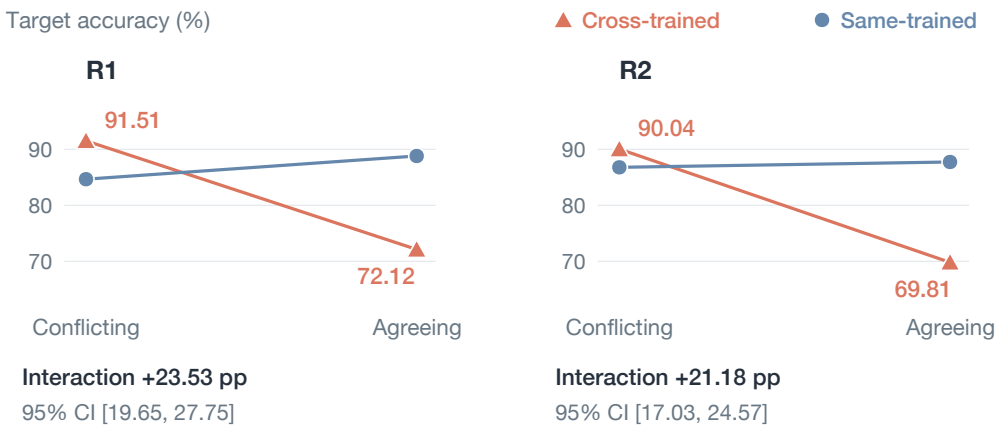


Figure 1: **Agreeing neighboring evidence can hurt.** Two independently trained ContractNLI pairs (783 units per condition; 38 equally weighted trios). Interactions are same-trained minus cross-trained accuracy changes from conflicting to agreeing companions, with paired 95% intervals; conditions use different companion contracts. All endpoint values appear in Appendix Table 9.

The controlled results connect this training choice to later use. Interleaving the same occurrences and partners moves the ProofWriter singleton contrasts near zero. Holding the shared-block checkpoints fixed, paired requests change their ordering: on theories absent from both SFT corpora, the singleton-selected 4B checkpoints trail alternatives by more than ten points. At 8B, reorganizing paired requests collapses rather than reverses the gap. In ContractNLI, weakening the designed disagreement raises training agreement to about 43% and attenuates both signatures predicted before new training, with retained paired-task competence and condition-specific costs. A later check over new requests on already-opened test documents shows that the same physical neighbors' relative effect depends on the target question in the original checkpoints and disappears in the repaired ones (Section 4).

Prior work already shows that sequence composition over the same documents can change downstream context use, and that matching exemplar-level supervision while changing which tasks share a sequence can teach different strategies (Zhao et al., 2024; Lee et al., 2025). We establish the consequences for complete SFT judgments under occurrence- and update-matched partner rewiring, then test a signed prediction from the designed pairing relation on natural contract text.

2 Sequence Pairing and Matched Design

Let an atomic supervision occurrence be $a_i = (x_i, y_i)$, where x_i is one complete problem input and y_i its target response. Repeated copies of the same text remain distinct occurrences, so multiplicity is part of the fixed multiset V. A partner map π assigns each occurrence to one ordered two-item row. A serialization map σ converts a row into a token sequence $z_r = \sigma(r)$ and designates the answer-token positions $\mathcal{Y}_r(\sigma)$ covered by response-only loss. Training under this realized construction minimizes

$$\mathcal{L}(\theta; \pi, \sigma) = -\sum_{r \in \pi} \sum_{t \in \mathcal{Y}_r(\sigma)} \log p_\theta(z_{r,t} \mid z_{r,<t}) \,. \tag{1}$$

At fixed serialization, matching atomic responses fixes their supervised tokens and multiplicities, but not the prefixes $z_{r,<t}$ under which they are predicted. Changing π can therefore change the conditional training problem. Changing σ can change both serialized targets and prefixes while preserving atomic labels and partners.

Partners and layouts. The shared-answer-block layout writes Q_1, Q_2, A_1, A_2, placing both problems before both supervised answers. The interleaved layout writes Q_1, A_1, Q_2, A_2, with loss on both assistant turns and user turns masked. Its first answer has an arm-identical prefix because Q_2 has not appeared; the second-position occurrence follows its reassigned partner and its own question. The indexed two-answer response also becomes two bare-label assistant turns: this is a complete layout change, not just a prefix edit.

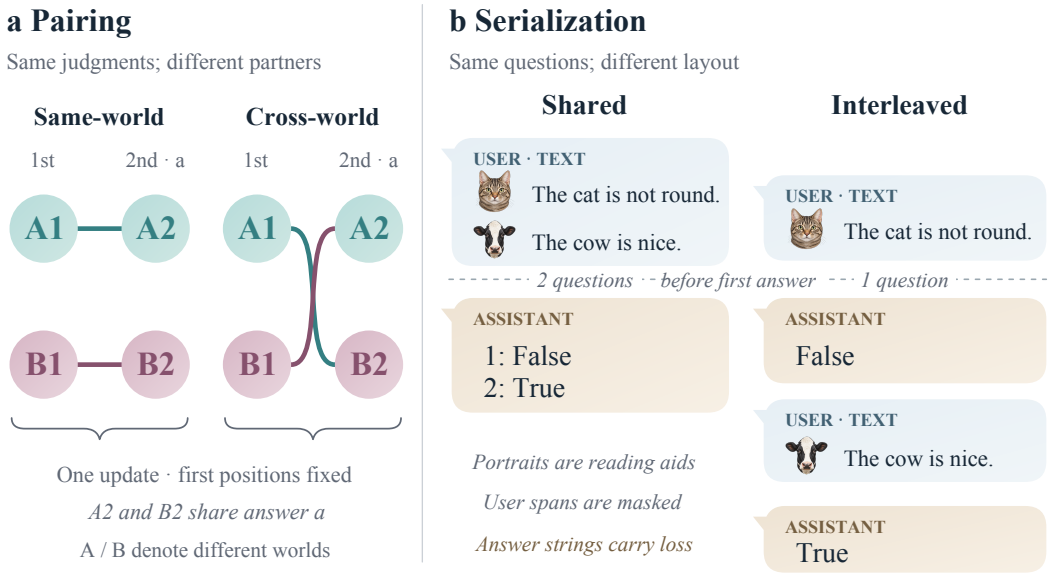


Figure 2: **Equal judgments leave pairing and prefixes unspecified.** (a) Within-update rewiring preserves occurrences, position roles and rowwise answers; the schematic second-position answers share label a. (b) A persisted cat/cow row under shared-block and interleaved training; theory text is elided and the shared answer retains its two-line form.

In same-world rows, two distinct conditions refer to one world. Cross-world rewiring fixes every first-position occurrence and bijectively reassigns second-position occurrences within each optimizer update and second-label class, requiring different worlds in every resulting row. Every occurrence retains its position role and update membership. Fixed first positions and second labels preserve ordered label pairs and, within a layout, exact rowwise response strings. Aggregate prompt, sequence and supervised-token exposure, optimization budget and checkpoint rule are matched within each comparison. Row-local theory repetition, wording and length composition change with the partner: these are consequences of the implemented intervention, not pointwise-matched quantities.

Training realizations. ProofWriter supplies deterministic labels for complete world–condition problems. Each arm contains the same 20,000 occurrence-indexed judgments in 10,000 two-question rows, assigned by a persisted permutation to 250 updates of 40 rows. We use Qwen3 (Yang et al., 2025) no-thinking, response-only full SFT and evaluate only checkpoint 250. Within a realization, arms share training RNG, row order, sampler permutation and optimization. The two 4B shared-block realizations, R1 and R2, differ only in training RNG, not data order. Their interleaved counterparts I1 and I2 train new checkpoints with the aligned occurrences, partners, update memberships, RNGs and singleton readout. Token totals are matched within, not across, layouts. R1-8B and R2-8B use the same rows and schedule at Qwen3-8B and are evaluated only on the separate unseen-theory paired population (Section 3.2). Exact census, recipes and matching audits appear in Appendices A and C.

ContractNLI applies fixed-update rewiring to complete contract–hypothesis judgments: the same 20,000 judgments in 10,000 rows receive 250 updates, paired within or across contracts. Within each update, arms match target occurrences, companion judgments, position roles, two-label responses and token exposure. Its independent second pair changes the model seed and update-block order, then uses the same evaluation matrices. The natural-text readouts are introduced in Section 4.

Consumers and endpoints. A *consumer* is the complete request, inference and scoring procedure. Request organization—how targets and companions are grouped—is one choice within it. The singleton consumer presents one of 192 held-out world–condition cells, requests exactly one label under an eight-token cap, and decodes once greedily with no thinking. This split holds out target cells, not whole worlds. Singleton exact correctness requires the complete trimmed response to equal the gold label, independent of finish status. Paired consumers require a two-line schema and score either both answers (exact-pair) or one designated target slot; their populations are introduced with the

results (Appendix B). Within a consumer, arm contrasts are cross-world minus same-world training. Reported 95% intervals measure item- or cluster-sampling uncertainty conditional on the retained checkpoints, not variation across training runs. The partner-placement and RACE probes supply scope boundaries (Appendices A.1 and E).

3 Controlled Evidence on ProofWriter

3.1 Partner rewiring changes delivered answers

Shared-answer-block R1 scores 27.08% cross-world versus 89.58% same-world on singleton exact correctness, a -62.50-point contrast ($[-69.27, -55.21]$). R2 shows the same direction: 0.00% versus 47.40%, a -47.40-point contrast ($[-54.69, -40.63]$). Within every optimizer update, the two arms contain the same atomic occurrences and response strings; only the second-position partner changes. The cross-world checkpoints fail to return exactly one allowed label on $140/192$ R1 and $192/192$ R2 singleton requests. Reading only the first standalone allowed label in those same saved outputs changes the cross-minus-same contrasts to $+5.21$ and $+44.27$ points: strict correctness measures the complete delivered response, so its gap cannot by itself be read as a form-independent reasoning deficit; first-label reading is only a permissive diagnostic (Appendix D). The same fixed checkpoints reverse this ordering on a high-compliance two-line paired interface; before turning to that comparison, we measure the role of training layout.

Training the same partners in another layout. I1 and I2 train new checkpoints from the R1- and R2-aligned occurrences and partner maps as interleaved question–answer turns. Under the same singleton consumer, the cross-minus-same contrasts are -0.52 and -1.04 points. All four arms exceed 92% exact correctness, and all 768 responses are complete allowed labels (Figure 3; Table 4 in Appendix C). The change bundles target placement, response units, turn structure and prefixes. Near-zero contrasts are local no-detected-contrast results for this consumer, not equivalence between layouts. The appendix also reports the post-hoc complete-layout changes on the same 192 targets.

An independent four-message control crosses partner assignment with whether Problem 2 is visible before the first supervised answer. Its two interactions lie inside the prespecified ± 10-point band (Appendix A.1), so this narrow placement change alone does not reproduce the complete-layout movement.

High singleton label compliance under interleaving does not imply compatibility with the shared two-line paired request: all four interleaved checkpoints return one bare label to that request (Appendix E).

3.2 Request organization changes checkpoint comparison and selection

Paired use reverses the checkpoint preference. Holding the four shared-block checkpoints fixed, we present the same 192 targets as 96 two-question requests whose problems come from different worlds. Across the four checkpoints, 383 of 384 responses satisfy the required two-line schema. A target is scored correct only when the full response has that schema and its own slot is correct. This target-slot score favors cross-world training by $+9.90$ points in R1 and $+17.71$ points in R2 (Figure 3). The prospective exact-pair endpoint, requiring both answers to be correct, likewise favors cross-world training by $+16.67$ and $+25.00$ points (Table 3 in Appendix C). Under this high-compliance paired interface, the opposite target-correctness ordering shows that the singleton preference does not carry across requests.

The contrast returns to the zero neighborhood with a native companion, and the same organization dependence remains when that native companion is absent from both SFT arms (Appendix E); exact reuse of an SFT-exposed companion is therefore unnecessary in these comparisons.

The checkpoint difference follows request organization at both scales. We next hold the paired interface and content pool fixed while varying request organization: whether each target keeps its native companion or receives a swapped one. Here native and crossed describe request organization, whereas same-world and cross-world describe the training arms. The evaluation uses 192 complete theories absent from both SFT corpora. Each supplies a target and a different-condition companion; swapping companions within 96 two-theory clusters turns native requests into crossed requests

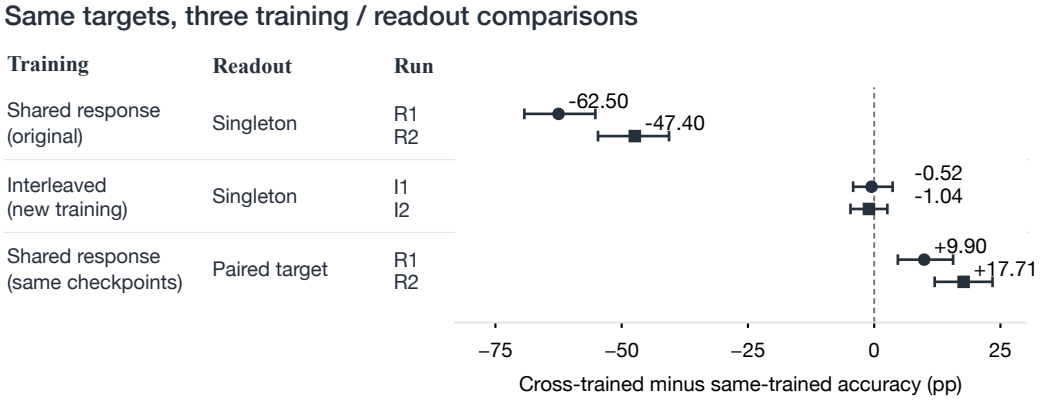


Figure 3: **Pairing contrasts depend on layout and readout.** Cross-minus-same accuracy on the same 192 targets, with paired 95% intervals. The first and third groups reuse R1/R2; I1/I2 are separately trained with interleaved turns. Singleton correctness requires the complete allowed label to be correct; paired target correctness requires the two-line schema and a correct target slot.

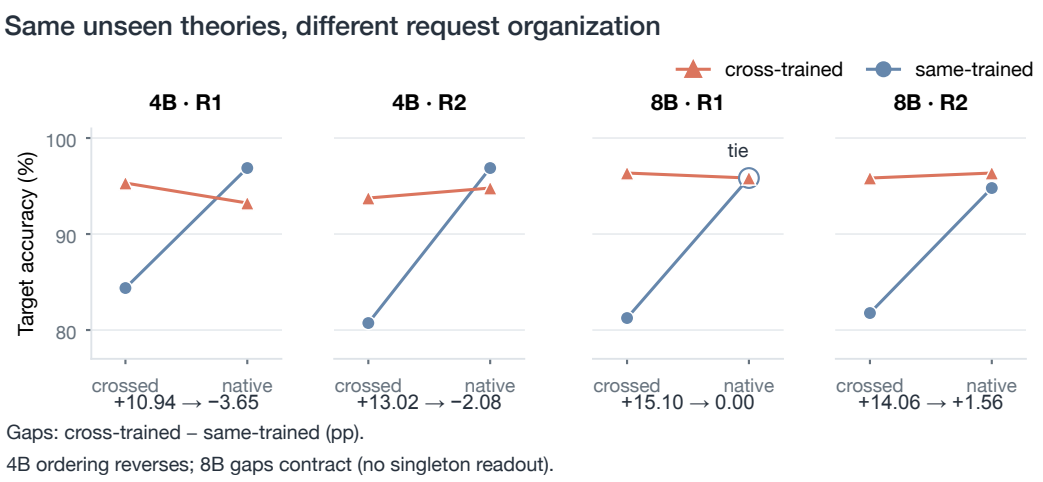


Figure 4: **Request organization changes the checkpoint comparison.** Both scales use the same 192 theories absent from both SFT corpora. Native requests reverse the 4B point-estimate ordering and shrink the 8B gaps to 0.00 and 1.56 points. Organizations preserve target and companion occurrences but change their assignment and within-request repetition. Gaps are cross-trained minus same-trained target-slot accuracy in percentage points. Contrast intervals and all 16 cells appear in Appendix Table 7.

while preserving the target and companion multisets. Every checkpoint-by-organization cell is at least 191/192 schema-valid. Reorganization moves the arms asymmetrically (Figure 4). Native organization raises the same-world-trained checkpoints by 12.50–16.15 points across both scales, while the cross-world-trained checkpoints move by at most 2.08 points in magnitude. At 4B, crossed-request gaps of +10.94 and +13.02 points become negative under native organization, reversing the point-estimate ordering. At 8B, +15.10 and +14.06-point gaps contract to a tie and a 1.56-point residual. The result is an organization-sensitive checkpoint difference in all four fixed realizations, with reversal only at 4B.

Retrospectively, singleton exact correctness would select the same-world checkpoint in both 4B realizations, yet those checkpoints trail on crossed paired requests over the separate SFT-unseen

theories. The 8B checkpoints have no singleton readout. Checkpoint selection should reflect the request organization expected at use time.

4 A Designed Training Relation Predicts Natural-Text Information Use

We now compare the within- and across-contract training arms on ContractNLI (Koreeda & Manning, 2021), repeating the same readouts for the independent second training pair. An agreeing or conflicting companion is defined by its official label on the target hypothesis, not by the answer to the companion question.

Target-consistent neighbors can hinder target judgments. On official ContractNLI test requests, changing from a conflicting to an agreeing companion lowers target accuracy for cross-contract-trained checkpoints by 19.39 and 20.23 points in two independent training realizations. In the first, accuracy falls from 91.51% to 72.12% despite the companion's agreement with the target answer. The corresponding same-contract-trained changes are $+4.14$ and $+0.95$ points; the latter has interval $[-1.62, +3.05]$. The same-minus-cross training-by-context interactions are $+23.53$ $[+19.65, +27.75]$ and $+21.18$ $[+17.03, +24.57]$ points. What reproduces is the interaction, not every within-checkpoint simple effect. All condition accuracies appear in Appendix Table 9.

Removing target-consistent evidence corrects errors. On agreeing requests, we delete either the companion evidence consistent with the target label or a length-matched neutral span, leaving the target contract, both hypotheses, response format, decoder, and scoring unchanged. Removing target-consistent evidence rather than neutral text raises target accuracy for the two cross-contract-trained checkpoints by 19.86 and 16.37 points. The same-contract-trained changes are -3.64 and 0.00 points. The corresponding same-minus-cross training-by-deletion interactions are -23.49 $[-38.55, -12.84]$ and -16.37 $[-25.17, -9.01]$ points. Of the 18 units corrected in the first realization, 16 were already wrong before either edit; in the second, all 15 corrections repair pre-existing errors and all 96 neutral-deletion predictions match the undeleted requests. In one frozen test-time case, contract 540 keeps its own return clause; deleting only the parallel clause from companion contract 77 changes the cross-trained target answer from `NotMentioned` to the correct `Entailment`, while the second answer stays fixed (Figure 5).

A designed relation yields a prospective prediction. The test-time comparisons above reveal a behavioral anomaly, not its cause. The frozen construction selected two different contracts that assign different official labels to a target hypothesis h while agreeing on a distinct companion hypothesis g. Cross rows pair the target contract's h judgment with the other contract's g judgment. Thus the companion contract assigns a different label to h in every original cross row: target-hypothesis agreement is exactly 0% by design, whereas same-contract rows necessarily agree on h. The companion's answer to h is absent from this row's two explicit answers, although reciprocal rows can supervise it elsewhere in the corpus. Agreement on g keeps the second-position response label unchanged under rewiring.

A model could learn to use the neighboring evidence as an oppositely signed cue. Weakening this training relation therefore gives a directional test of that account. We constructed one repaired cross-contract schedule for each training realization by changing which companion atoms were paired. The repair preserved every target occurrence, explicit judgment, row position, optimizer-update membership, per-update companion-atom multiset, token total, and the 250-update recipe. Agreement rose to 42.75% and 42.67%. These schedules weaken the relation rather than establish independence: companion identity, wording, length, and local layout still change with the reassignment. Before new training or model calls, we froze the existing official-test paired-use, source-agreement, and deletion readouts for both schedules, predicting an upward shift of at least five points in each source gap and a reduction of at least five points in each deletion contrast. The near-zero source gaps and exact-zero deletion contrasts reported below were not themselves prespecified values.

Re-pairing attenuates the predicted signatures. The two prospective tests support the relation explanation in this construction. The agreeing-minus-conflicting gaps move from about twenty points below zero to -0.18 and -0.33; the repaired target predictions differ between paired agreeing and

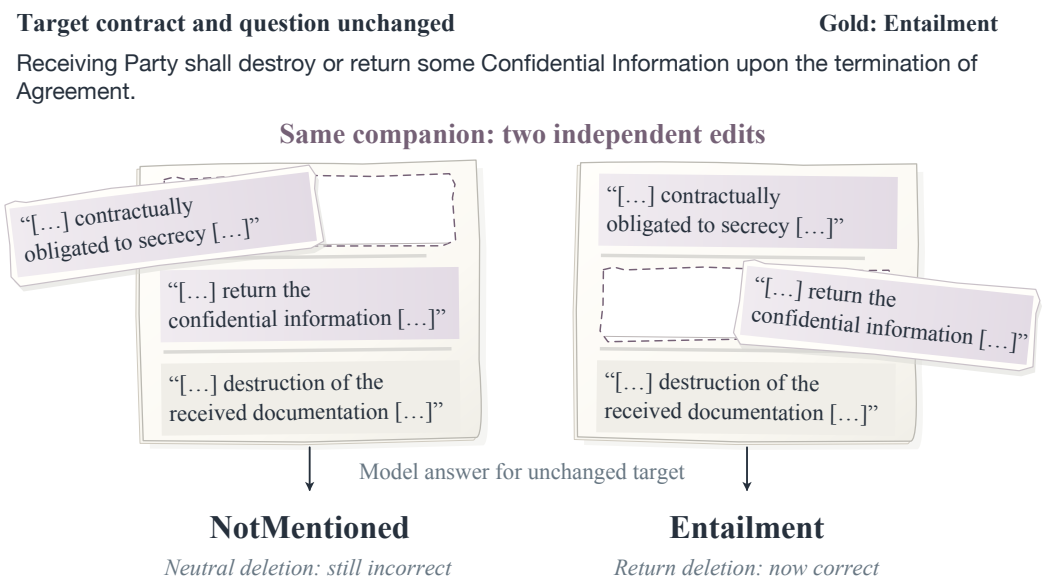


Figure 5: **Removing evidence from the neighbor can correct the target.** One illustrative case from the first cross-trained checkpoint: independent edits of the same companion leave the target contract and question fixed. The unedited target answer was also `NotMentioned`; answers beneath the edited documents show the contrast. Paper shapes and cutouts are schematic passage locators, not full deletion boundaries (complete spans and qualifications: Appendix E). Both edits retain the clause allowing destruction instead of return; the second answer stays `NotMentioned`. The same-trained checkpoint was already correct.

conflicting requests on only 4 and 5 of 783 targets. The deletion advantage disappears with it: both repaired contrasts are exactly zero, and each checkpoint gives identical target predictions under the two deletion requests on all 96 units (Figure 6; complete results in Appendix Table 8). These near-zero contrasts do not arise from general paired-task failure: the repaired checkpoints retain 84.70–85.66% target accuracy on document-bearing native and crossed paired requests, versus 44.81% when both documents are removed; all six paired-use cells are $732/732$ strict-schema-valid. This shows retained document-conditioned competence, not itemwise use of the non-target companion document.

The zero deletion contrasts accompany substantial improvements on neutral-deleted requests, while changes under target-evidence deletion are smaller and uncertain (Appendix E). Re-pairing also has costs: agreeing-condition accuracy rises by $+15.79$ and $+18.43$ points, whereas conflicting accuracy changes by -3.41 and -1.47 points; only the first decline has an interval wholly below zero. The prespecified criterion for useful-content improvement was not met. The repair supports the predicted attenuation without establishing uniformly better training.

The same neighbors have a question-dependent effect. After repair training and its main readout, we froze a further result-blind test on new request combinations over already-opened official-test documents. Each group fixes the target contract and gold label, two physical companion contracts, and the companion question, then crosses eligible target questions for which the agreeing companion changes sides. Across 38 disjoint document triples, the original cross-trained checkpoints show interactions of -17.32 and -22.25 points, with paired-triple intervals $[-23.75, -10.89]$ and $[-27.25, -17.42]$. A question-independent companion offset would cancel on this correctness-difference scale: the same neighbors' relative effect depends on what is asked, though the question's wording and evidence target change with it.

Each repaired checkpoint instead gives the same target answer with either companion on all 406 target-question units, yielding exactly zero interactions (Figure 6). Companion answers still change on 34/406 and 30/406 units, so the invariance is target-specific. Repair also makes 26/812 and 29/812 previously correct target cells wrong; the full transition breakdown is in Appendix E.

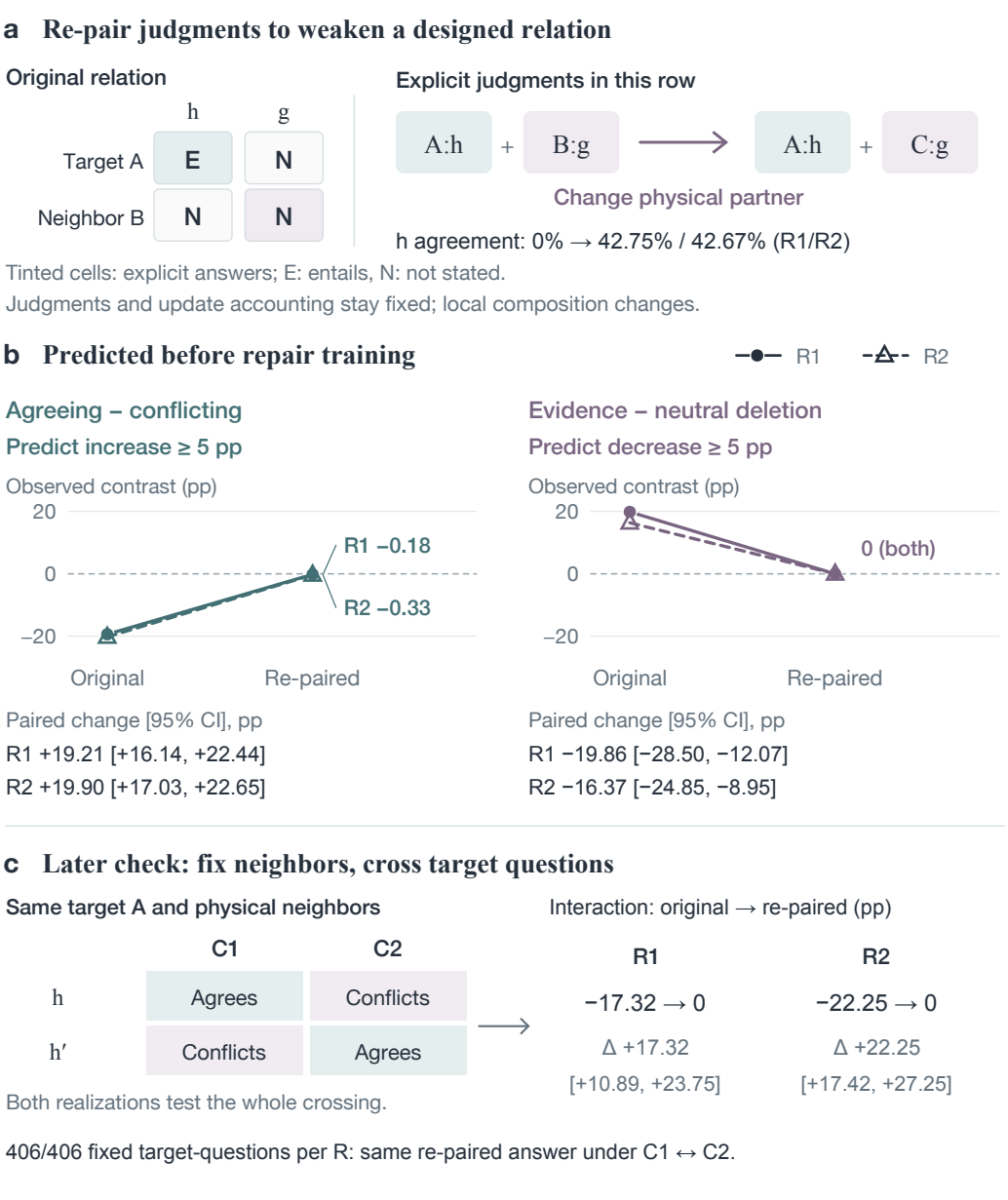


Figure 6: **A designed relation predicts changes in context use.** (a) Tinted cells are this row's explicit answers; re-pairing changes physical partners and local composition along with agreement on target question h (g: other question). (b) R1/R2 are independent training realizations. Source gap is agreeing-minus-conflicting target accuracy (38 trios); deletion contrast is target-evidence-minus-neutral deletion accuracy (33 trios, 96 targets). Lines compare matched original and newly trained re-paired checkpoints. Brackets throughout are paired 95% intervals of changes. Near-zero or zero endpoints were not predicted values. (c) The shared design fixes the target contract, gold label, physical neighbors and companion question while crossing target questions. Frozen after training, this check tests new requests on already-opened test documents (406 units). Attenuation coexists with condition-specific losses and retained document-conditioned competence, as detailed in the adjacent text.

5 Related Work

Packing, padding, shuffle granularity and optimizer-time order can change learning (Wang et al., 2025; Dong et al., 2025; Krell et al., 2022; Chen et al., 2026; Hu et al., 2026). With the same pre-

training documents, Zhao et al. change sequence composition and observe differences in distraction and downstream context use (Zhao et al., 2024). Related work varies target–context relatedness, demonstrations or knowledge beside a target (Malu et al., 2026; Lee et al., 2026). Lee et al. train on every exemplar, match exemplar-level inputs, outputs and task loss weights across curricula, then test the resulting strategy by parameter mismatch (Lee et al., 2025). Matched examples and multiple loss-bearing judgments are therefore not distinctions on their own.

We hold complete judgment occurrences, answers and optimizer-time placement fixed while rewiring sequence partners, then follow the behavior through a training-layout change, fixed-checkpoint selection and signed evidence use on natural contract text. In ContractNLI, each judgment supplies its own problem: its gold answer does not require inferring a task function from earlier examples. The designed pairing relation yielded a prediction tested in two new training realizations. Attention-isolated packing is a different regime because our intervention requires one supervision unit to enter another's causal prefix.

Format studies distinguish task content from output convention and first-token probabilities from generated answers (Liang et al., 2024; Bunn et al., 2025; Wang et al., 2024). Our rewiring arms receive the same response format, ordered label pairs and response strings; differences in interface behavior are outcomes of the construction, not supplied arm differences.

6 Discussion, Limitations, and Conclusion

Complete SFT judgments can be learned in different surrounding histories. Our comparisons show why those histories matter beyond a training score: partner assignment changes delivered behavior, and a checkpoint preference can change with its downstream requests. Natural contract text supplies a sharper example. Evidence in a neighboring contract agrees with the target yet hinders these checkpoints; a relation deliberately imposed during training suggests a directional prediction that holds in two new training realizations. SFT studies with shared autoregressive histories should report and audit relations between sequence partners, and compare checkpoints under the requests expected at use time.

Limitations. The principal comparisons have two matched training realizations, testing repeatability rather than a seed-population effect or stable effect size. The principal evidence uses two scales in one model family, with all main trainings ending at 250 updates. Complete-layout and consumer comparisons change bundled treatments, not a single microscopic coordinate. Native two-turn history dependence is clear in one interleaved realization but uncertain in the other; additional scoring and reading-comprehension probes set scope boundaries (Appendix E).

In ContractNLI, re-pairing changes companion identity, wording, length and local layout along with agreement. A control intended to separate the relation from physical companion replacement failed its prespecified construction threshold and was not trained. Later question-cross requests reuse already-opened test documents and change question wording and evidence targets. The repair incurs condition-specific losses and adverse target transitions, and misses its prespecified useful-content criterion. These are construction-level behavioral findings, not an isolated internal mechanism or a generally better training rule.

The same complete judgments, paired differently, can teach different uses of neighboring information. The natural-text prediction and the checkpoint-selection consequence make this pairing choice worth treating as an explicit part of SFT design.

AI USE STATEMENT

Generative AI tools supported literature review, refinement of research ideas, mathematical and computational analysis, and manuscript preparation and polishing. The authors led the research, directed the use of these tools, and made the final scientific decisions. The authors take responsibility for the final content.

REFERENCES

Alec Bunn, Sarah Wiegreffe, and Ben Bogin. Fine-tune on the format: First improving multiple-choice evaluation for intermediate LLM checkpoints. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics*, pp. 511–521, 2025. URL https://aclanthology.org/2025.gem-1.46/.

Yanbing Chen, Ruilin Wang, Zihao Yang, Lavender Yao Jiang, and Eric Karl Oermann. Optimizing packing and shuffling strategies for enhanced performance in generative language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 1399–1416, 2026. doi: 10.18653/v1/2026.acl-srw.124. URL https://aclanthology.org/2026.acl-srw.124/.

Jiancheng Dong, Lei Jiang, Wei Jin, and Lu Cheng. Threshold filtering packing for supervised fine-tuning: Training related samples within packs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4422–4435, 2025. doi: 10.18653/v1/2025.naacl-long.226. URL https://aclanthology.org/2025.naacl-long.226/.

Xiaomeng Hu, Yixuan Tang, Haoze Li, Hao Chen, Qi Zhang, Zhanming Shen, Yiming Zhang, Haobo Wang, and Junbo Zhao. Fine-grained data ordering improves fine-tuning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20406–20418, 2026. doi: 10.18653/v1/2026.findings-acl.1021. URL https://aclanthology.org/2026.findings-acl.1021/.

Yuta Koreeda and Christopher D. Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1907–1919, 2021. doi: 10.18653/v1/2021.findings-emnlp.164. URL https://aclanthology.org/2021.findings-emnlp.164/.

Mario Michael Krell, Matej Kosec, Sergio P. Perez, and Andrew Fitzgibbon. Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance. *arXiv preprint arXiv:2107.02027*, 2022. doi: 10.48550/arXiv.2107.02027. URL https://arxiv.org/abs/2107.02027.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, 2017. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082/.

Hyunji Lee, Seunghyun Yoon, Yunjae Won, Hanseok Oh, Geewook Kim, Trung Bui, Franck Dernoncourt, Elias Stengel-Eskin, Mohit Bansal, and Minjoon Seo. Instruction tuning with and without context: Behavioral shifts and downstream impact. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6790–6810, 2026. doi: 10.18653/v1/2026.eacl-long.320. URL https://aclanthology.org/2026.eacl-long.320/.

Jin Hwa Lee, Andrew K. Lampinen, Aaditya K. Singh, and Andrew M. Saxe. Distinct computations emerge from compositional curricula in in-context learning, 2025. URL https://arxiv.org/abs/2506.13253v1. arXiv:2506.13253v1.

Shihao Liang, Runchu Tian, Kunlun Zhu, Yujia Qin, Huadong Wang, Xin Cong, Zhiyuan Liu, Xiaojiang Liu, and Maosong Sun. Exploring format consistency for instruction tuning. *arXiv preprint arXiv:2307.15504*, 2024. doi: 10.48550/arXiv.2307.15504. URL https://arxiv.org/abs/2307.15504.

Deeptanshu Malu, Deevyanshu Malu, Aditya Nemiwal, and Sunita Sarawagi. Training in-context and in-weights mixtures via contrastive context sampling. *arXiv preprint arXiv:2604.01601*, 2026. doi: 10.48550/arXiv.2604.01601. URL https://arxiv.org/abs/2604.01601.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, 2021. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

Shuhe Wang, Guoyin Wang, Yizhong Wang, Jiwei Li, Eduard Hovy, and Chen Guo. Packing analysis: Packing is more appropriate for large models or datasets in supervised fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4953–4967, 2025. doi: 10.18653/v1/2025.findings-acl.256. URL https://aclanthology.org/2025.findings-acl.256/.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, 2024. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441/.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Yu Zhao, Yuanbin Qu, Konrad Staniszewski, Szymon Tworkowski, Wei Liu, Piotr Miłoś, Yuxiang Wu, and Pasquale Minervini. Analysing the impact of sequence composition on language model pre-training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7897–7912, 2024. doi: 10.18653/v1/2024.acl-long.427. URL https://aclanthology.org/2024.acl-long.427/.

Table 1: **Realized ProofWriter matching audit for R1/R2.** Counts are per arm under the shared-answer-block layout. The aggregate token totals are identical, while local row composition is not claimed to match pointwise.

Quantity	Same-world	Cross-world
Training rows	10,000	10,000
Atomic occurrences	20,000	20,000
Ordered label pairs	identical	
Response-string marginals	identical	
Prompt tokens	2,740,827	2,740,827
Templated sequence tokens	2,889,694	2,889,694
Supervised response tokens	90,000	90,000
Dropped / truncated rows	0 / 0	0 / 0
Mean sequence length	288.9694	288.9694
Median / 95th percentile	281 / 442	287 / 393
Maximum sequence length	588	506
Training cutoff	768	768

A Construction, Serialization, and Matching Audit

Occurrence-indexed supervision. The ProofWriter source contains 4,992 degree-3 worlds and 9,984 world–condition cells. An atomic supervision occurrence contains one complete world, one condition, and its deterministic label in $\{\texttt{True}, \texttt{False}, \texttt{Unknown}\}$. Occurrences, rather than deduplicated texts, are the nodes of the intervention: repeated draws of the same cell remain distinct nodes, and their multiplicities are matched item by item. Each arm contains 10,000 ordered two-question rows, hence 20,000 atomic occurrences from the same occurrence-indexed multiset.

Partner reassignment within fixed updates. In a same-world row, two different conditions are posed against one world. Before training, a persisted permutation assigns the 10,000 row indices to 250 optimizer-update blocks of 40 rows. Within each block, every first-position occurrence remains fixed, and second-position occurrences are bijectively reassigned only within their second-label class; each resulting cross-world row must contain problems from different worlds. No occurrence is added, removed, relabeled, or moved to another optimizer update. At a fixed serialization, each row retains its ordered answer pair and response string. The intervention changes which second-position occurrence shares a causal training sequence with each fixed first-position occurrence.

Shared-answer-block serialization and realized accounting. Each R1/R2 row contains two complete world–condition prompts followed by the shared response schema:

```
1: <label>
2: <label>
```

The same-world arm repeats the common theory inside the two complete problem blocks, whereas a cross-world row contains two different theories. This serialization is part of the realized sequence construction: both complete problem blocks precede both answer lines, so reassigning the second problem changes the realized prefix under which both supervised answers are predicted even though the first problem occurrence and both target strings remain fixed rowwise. Both arms use the same native Qwen3 chat template and response-only loss mask, with no additional sequence packing. Table 1 reports the accounting recomputed from the persisted corpora after template preprocessing. These totals describe the shared-answer-block layout; interleaved totals are audited within layout rather than claimed equal across layouts.

Rewiring changes which prompt contents, prefix histories, and prompt lengths co-occur within an individual row. The two sequence-length histograms overlap by 74.36% (total-variation distance 25.64%); row-level length and prefix composition are therefore not matched pointwise. Both observed maxima remain below the 768-token cutoff, so this difference induces neither truncation nor unequal aggregate exposure.

Table 2: **Partner-placement interaction.** Values are cross-world minus same-world singleton exact correctness in percentage points. The final column is partner-before-A_1 minus partner-after-A_1.

Training seed	Partner after A_1	Partner before A_1	Interaction [95% CI]
1991603944	-2.60	$+0.52$	$+3.13$ $[-3.13, +9.90]$
1353929150	$+4.17$	$+2.60$	-1.56 $[-8.33, +5.21]$

Interleaved serialization audit. I1 and I2 re-serialize the identical R1- and R2-aligned rows as user Q_1, assistant A_1, user Q_2, assistant A_2 under the same native Qwen3 template, with loss on both assistant turns and user turns masked. Row counts, occurrences, multiplicities, slot roles, atomic labels, partner maps, epoch-0 sampler permutations, and per-occurrence update membership are inherited from the corresponding shared-answer-block realization. The serialized targets are not inherited: the shared block's single indexed two-line response becomes two bare-label assistant turns. All four arms retain 10,000 rows with zero dropping and truncation and equal trained-token exposure within each pair. Because the turn structure differs, I1/I2 are audited within layout and their templated sequence totals are not compared with Table 1.

A.1 A PARTNER-PLACEMENT CONTROL

The shared-block/interleaved bridge changes a complete training layout. To probe one candidate coordinate, a separate factorial crosses same- versus cross-world within-update partner assignment with two four-message layouts. The *partner-before-*A_1 layout (persisted implementation label: `preposed`) begins with a user message containing the common two-problem introduction, the complete Problem 1 and Problem 2 blocks, and the instruction to answer Problem 1. The assistant returns bare-label A_1; a second user message contains only the instruction to answer Problem 2; and the assistant returns bare-label A_2. The *partner-after-*A_1 layout (persisted implementation label: `interleaved`) begins with a user message containing the common introduction, the complete Problem 1 block, and its answer instruction. The assistant returns bare-label A_1; the second user message contains the complete Problem 2 block and its answer instruction; and the assistant returns bare-label A_2. Thus both role sequences are user, assistant, user, assistant, both assistant turns carry loss, and user turns are masked. These cells are separately trained artifacts under their own audited native enabled-thinking render, not R1/R2 or I1/I2.

Within a seed, the four cells match atomic occurrences, the partner map for each topology across layouts, update-level exposure, supervised answer sequence, optimizer schedule, training recipe, and checkpoint-250 rule. Each cell contains 10,000 rows and receives exactly 250 updates. All 80,000 supervised target-prefix checks match the native enabled-thinking serving render. Each of the eight checkpoints answers the same immutable 192-target bare-label population once with greedy decoding and `max_tokens`=8. All 1,536 planned responses arrive in exactly 1,536 attempts. One length-capped response remains in the denominator without regeneration or parser repair; all cells pass the frozen response floors.

For seed s and layout ℓ, let $\Delta_{s,\ell} = A^{\text{cross}}_{s,\ell} - A^{\text{same}}_{s,\ell}$. The primary estimand is $J_s = \Delta_{s,\text{before}} - \Delta_{s,\text{after}}$. Its interval uses 10,000 paired block-stratified bootstrap resamples with seed 2086741903 and is judged against the prespecified ± 10-point meaningful band.

Both interaction intervals lie wholly within the prespecified band. The result bounds the mean partner-placement interaction in this independent four-message construction; it does not decompose the shared-block/interleaved comparison.

B LAYOUTS, READOUT POPULATIONS, AND SCORING

Training layouts. The R1/R2 and I1/I2 layouts serialize the complete world–query input for each occurrence. The shared-answer-block layout places two problem blocks before one assistant response containing two indexed answer lines. The interleaved layout instead produces user Q_1, assistant A_1, user Q_2, assistant A_2, with one bare label in each assistant turn. Loss covers both assistant turns and masks the user turns. The layouts preserve atomic labels and partner assignments but use different serialized target strings and turn boundaries.

Scoring rules. Singleton exact correctness trims surrounding whitespace and accepts only a complete response equal to the gold label. The permissive diagnostic selects the first standalone allowed label in the saved response. Paired exact correctness requires two indexed lines in order and both labels correct; target-slot correctness applies the same complete-response schema check but scores only the designated target in its original slot. No parser repairs a missing answer, reorders labels, filters rows, or triggers regeneration.

Companion-request coverage. The cross-world-companion population consists of 96 two-question requests formed from the same 192 singleton targets, with 32 pairs in each of three source blocks. Each target appears once in its assigned slot, and both slots are held-out targets. The identical 96 requests are issued to each of the four R1/R2 checkpoints and, separately, to each of the four I1/I2 checkpoints, yielding 384 responses from each checkpoint group. Two same-world-companion controls instead issue 192 target-specific requests to each of the four shared-block checkpoints. Every target keeps its original slot and receives a different-condition companion from its native world, drawn either from training-exposed cells under the frozen label-blind rule or from unique cells absent from both SFT training arms and the held-out target set. Each condition yields 768 responses; only the designated target slot contributes to either control's primary endpoint.

Unseen-theory paired population. The transfer population contains 192 unique ProofWriter theories absent from both SFT training arms and from the structured design, request, and evaluation manifests predating its original selection; selection does not consult labels or model outputs. The population and both organizations are reused unchanged for the Qwen3-8B realizations. This exclusion does not cover base-model pretraining, for which exposure is unknown. Following the original 192-slot composition template, 187 theories exactly match `config`, `maxD`, `NFact`, `NRule`, and the target structural condition. Five deterministic fallbacks retain `config`, `maxD`, and the exact target condition. Each theory supplies one target and one companion with a different structural condition. The theories form 96 deterministic two-theory clusters. In the native organization, each target receives its native companion; in the crossed organization, the two companions are exchanged within each cluster. Both organizations therefore contain exactly the same 192 targets and 192 companion occurrences at the same multiplicities. Each of the four Qwen3-4B shared-answer-block checkpoints and each of the four Qwen3-8B checkpoints answers both organizations once, and only the target slot contributes to the endpoint.

Persisted examples. One training row at canonical index 5663, optimizer update 85, and within-update position 34 contains the queries "The cat is not round." and "The cow is nice.", with labels `False` and `True`. With unchanged theory text elided, its shared-block target is `1: False / 2: True`; the aligned interleaved row places the loss-bearing answer `False` after the first problem and `True` after the second. This persisted row is used in Figure 2.

A persisted paired-readout example comes from canonical index 33 of the R1 readout. The first query is "Gary is green."; the theory gives Gary as white and derives cold, rough, then green, so its label is `True`. The companion query, "The tiger is not green.", has gold label `False`. The R1 cross-world-trained checkpoint returns `1: True / 2: False`, whereas the same-world-trained checkpoint returns `1: Unknown / 2: False`; both satisfy the requested two-line schema. This instance illustrates delivered target correctness. The opposite singleton ordering is established by the aggregate comparison, not inferred from this one example.

C Training, Checkpoint, Serving, and Uncertainty

Models and optimization. The R1/R2 and I1/I2 arms use Qwen3-4B (Yang et al., 2025) at revision `1cfa9a7208912126459214e8b04321603b3df60c`, with no-thinking, response-only full SFT without additional packing. Training uses four GPUs, micro-batch size 1 per GPU, gradient accumulation 10, and global batch size 40. Optimization uses AdamW at learning rate 2×10^{-5}, cosine decay with 10 warmup updates, mixed BF16, gradient checkpointing, FlashAttention 2, and DeepSpeed ZeRO-3. Every arm receives exactly 250 optimizer updates. Only checkpoint-250 is evaluated; 10,000 rows at global batch 40 make this a nominal single pass over pair instances.

Table 3: **The same Qwen3-4B checkpoint pairs trade places across complete consumers.** Rows report cross-world minus same-world training contrasts in percentage points with paired 95% intervals. Exact-pair requires both answers to be correct; target-slot scores a designated question under the full two-line schema. Table 7 reports the separate theory-level population used for the cross-scale comparison.

Consumer / endpoint	Realization 1	Realization 2
Singleton / exact correctness	-62.50 $[-69.27, -55.21]$	-47.40 $[-54.69, -40.63]$
Cross-world companion / exact-pair	$+16.67$ $[+6.25, +27.08]$	$+25.00$ $[+15.63, +34.38]$
Cross-world companion / target-slot	$+9.90$ $[+4.69, +15.63]$	$+17.71$ $[+11.98, +23.44]$
SFT-exposed native companion / target-slot	-1.56 $[-5.73, +2.08]$	-4.17 $[-8.33, 0.00]$

Qwen3-8B fixed-update realizations. R1-8B and R2-8B use Qwen3-8B at revision `b968826d9c46dd6066d109eabc6255188de91218` under the same no-thinking, response-only full-SFT optimization settings and checkpoint rule. They reuse the R1/R2 construction's 10,000 physical rows, physical row order, data seed 1991603944, native sampler permutation, and 250 effective optimizer-update blocks. R1-8B uses training RNG 1991603944 and R2-8B uses 1353929150; within each realization the arms differ only in second-position partner assignment. The four checkpoints are evaluated only on the unseen-theory paired population.

Frozen singleton readout. The original singleton population contains 192 held-out degree-3 world–condition target cells: 64 from each of three predefined, mutually disjoint source blocks, with no target-cell overlap with training or earlier readout populations. Other conditions from the same source theories may occur in training; the split does not hold out entire worlds. Every R1/R2 or I1/I2 checkpoint answers the identical 192 singleton prompts once under greedy no-thinking decoding with `max_tokens`=8. There is no output-dependent filtering, regeneration, or retry.

The R1/R2 study uses the same Qwen3-4B revision, optimization recipe, update count, checkpoint-250 rule, and realized data schedule in two training-RNG realizations. R1 uses training RNG 1991603944. R2 uses training RNG 1353929150 while retaining data seed 1991603944, the physical row order, and every persisted 40-row update block. Thus the second realization is not a second data-order draw. I1 and I2 reuse the corresponding R1 and R2 training RNGs, while both retain data seed 1991603944, the paired persisted row order, and update blocks; they differ from their aligned shared-answer-block realization only in serialization.

For the R1/R2 and I1/I2 studies, all 1,536 singleton, 384 shared-block cross-world-companion, 384 interleaved cross-world-companion, 768 training-exposed same-world-companion, and 768 training-unseen same-world-companion requests return once; every reported rate uses its complete frozen denominator. The singleton paired-world intervals use 10,000 resamples with the three-block stratification; cross-world-companion intervals use the corresponding frozen paired-row resampling. Same-world-companion target-slot intervals use 10,000 paired resamples over six source-block–by–target-slot strata with seed 2042092931. Training-unseen same-world-companion intervals use the same six strata with seed 3190914217. The separate Qwen3-4B unseen-theory readout issues 1,536 requests to the four shared-answer-block checkpoints, all returned once. Its intervals use 10,000 paired bootstrap replicates over the 96 complete mutual-swap clusters with seed 2533350915, applying every sampled cluster to all eight fixed prediction vectors. The identical paired population is read separately by the four Qwen3-8B checkpoints below (Appendix E).

Post-hoc singleton-to-paired contrast change. For the post-hoc fixed-artifact interaction, paired request i binds singleton targets $2i$ and $2i+1$ in their original slots, and the 96 pair clusters retain their three 32-pair source blocks. Singleton exact correctness and companion target-slot exact correctness are computed under their existing parsers; the prospective exact-pair endpoint is unchanged. Within each bootstrap replicate, pair indices are sampled with replacement inside each source block and both slots of every sampled pair are retained. The analysis uses 10,000 resamples with seed 2042092931. The cross-minus-same training contrast increases from singleton to cross-world-companion target-slot use by $+72.40$ points ($[+63.54, +81.25]$) in R1 and $+65.10$ points ($[+54.69, +75.52]$) in R2. This descriptive comparison aligns the original 192 target identities across two complete consumer contracts and is not a prospective endpoint.

Table 4: **Matched reserialization changes the singleton contrast.** Singleton exact correctness (%) and paired 95% cross-minus-same intervals. Each R/I pair retains occurrences, partner maps, optimizer-update placement, training RNG, and evaluation, but retrains under a different complete layout. Post-hoc changes between layout contrasts are reported separately.

Pair	Serialization	Cross	Same	Cross−same
R1	Q_1, Q_2, A_1, A_2	27.08	89.58	-62.50 $[-69.27, -55.21]$
I1	Q_1, A_1, Q_2, A_2	93.75	94.27	-0.52 $[-4.17, +3.65]$
R2	Q_1, Q_2, A_1, A_2	0.00	47.40	-47.40 $[-54.69, -40.63]$
I2	Q_1, A_1, Q_2, A_2	92.71	93.75	-1.04 $[-4.69, +2.60]$

Post-hoc complete-layout contrast change. For realization s, let A denote singleton exact correctness on the same 192 target identities and define

$$B_s = \left(A^{\text{cross}}_{I_s} - A^{\text{same}}_{I_s}\right) - \left(A^{\text{cross}}_{R_s} - A^{\text{same}}_{R_s}\right).$$

The analysis jointly aligns the eight immutable singleton prediction vectors for R1/R2 and I1/I2. It uses 10,000 paired block-stratified bootstrap resamples with seed 2086741903: the three 64-target source blocks are resampled separately, and the same sampled identities are applied to all eight vectors in each replicate. Each interval describes held-out-item uncertainty for its matched quartet of fixed checkpoints and targets. It is not a prospective layout factorial, a training-seed population interval, or an estimate for a single serialization coordinate. The resulting changes are $B_1 = +61.98$ points ($[+54.69, +69.27]$) and $B_2 = +46.35$ points ($[+38.02, +54.69]$).

RACE model revisions and cross-family realization. The Qwen RACE realization uses Qwen3-8B at revision `b968826d9c46dd6066d109eabc6255188de91218`. The cross-family realization uses `meta-llama/Llama-3.1-8B-Instruct` at revision `0e9e39f249a16976918f6564b8830bc894c89659`. It reuses the frozen RACE corpora, seed/data seed 1662733758, persisted update schedule, full-SFT recipe, global batch 40, 250 updates, and checkpoint-250 rule of the Qwen realization. Tokenization under the pinned Llama3 template retains all 10,000 rows in each arm without truncation. Both checkpoints are read once on the complete 4,887-question development population under the same single-question prompt, eight-token cap, parser, and paired passage-cluster bootstrap used for Qwen.

D Fixed-Update Readout Audit

Response forms and first-label scoring. In R1, the cross-world checkpoint produces 52 complete single-label responses, 47 multiple-label responses, and 93 responses containing one label plus extra text; all 52 exact-label outputs stop normally. The same-world checkpoint returns an exact label on all 192 requests, with 191 normal stops. In R2, all 192 cross-world responses contain multiple standalone labels and terminate at the generation cap. Across I1 and I2, all 768 responses are complete allowed labels; no response contains multiple labels or reaches the generation cap, and first standalone labels coincide with exact labels. The permissive diagnostic reads the first standalone allowed label anywhere in each saved response. Its cross-minus-same correctness contrasts are $+5.21$ points ($[+1.04, +9.38]$) in R1 and $+44.27$ points ($[+36.46, +52.08]$) in R2. These readings describe labels present in the generations; they do not replace the exact-primary endpoint.

Fixed-label ranking diagnostic. These eight R1/R2 and I1/I2 checkpoints are also read once without text generation by a fixed-label ranking consumer. It renders each persisted singleton instruction through the checkpoint's native chat template at the assistant position, computes `log_softmax` over the full next-token vocabulary, records the qualified single-token candidates `True`, `False`, and `Unknown`, and takes their unique within-set argmax; exact ties count as wrong. Across the resulting 1,536 checkpoint–target rows, the primary is relative-ranking accuracy. Intervals use 10,000 paired resamples within the three 64-row source blocks with seed 618502941. Candidate-set mass is the post-hoc sum of the three exponentiated full-vocabulary log probabilities and changes neither prediction nor endpoint.

Table 5 reports the complete conditional ranking contrasts, arm-prediction disagreements, and candidate-set coverage.

Table 5: **Fixed-label ranking diagnostics.** Accuracy values are cross/same percentages followed by their paired cross-minus-same 95% interval. Disagreement counts compare the two arm predictions on the same 192 targets. Candidate mass is the median total next-token probability assigned to the three fixed labels; it is descriptive and does not enter the ranking.

Realization	Accuracy, cross / same (contrast)	Disagreement	Median mass, cross / same
R1 shared block	91.15 / 90.10 ($+1.04$ $[-2.60, +4.69]$)	14/192	10.692% / 83.921%
R2 shared block	84.38 / 49.48 ($+34.90$ $[+27.60, +42.19]$)	85/192	0.055% / 85.997%
I1 interleaved	94.27 / 94.27 (0.00 $[-4.17, +4.17]$)	16/192	99.993% / 99.990%
I2 interleaved	91.67 / 93.75 (-2.08 $[-5.73, +1.04]$)	13/192	99.994% / 99.989%

In R2 cross, 174 of 192 rows assign less than 1% total probability to the three labels and all 192 assign less than 50%; the full-vocabulary top token was not recorded. The $+34.90$-point accuracy contrast therefore remains valid only as relative ranking conditional on the specified set. It neither removes answer emission nor establishes form-independent semantic recovery.

E Consumer, Construction, and Task Scope Protocols

E.1 ProofWriter request organization

Interleaved paired-contract scope. The I1 and I2 checkpoints also answer the 96 cross-world-companion pair requests used for their aligned shared-block realizations. Every arm scores 0/96 on schema-valid exact-pair correctness, while all 384 responses contain exactly one bare allowed label, more often the requested second label. This is a common response-contract failure under the paired request; the paired-use question under that shared two-line response contract is therefore unadjudicated for the interleaved layout.

A separate format-matched readout uses native user–assistant–user–assistant history: each checkpoint first answers a companion and then answers a target after either same- or cross-world history. Across 192 new targets in 96 mutual-swap clusters, the training-by-history interaction is $+4.69$ points ($[-1.56, +10.94]$) for I1 and $+10.94$ ($[+6.77, +15.10]$) for I2. The first misses the prespecified $+5$-point magnitude threshold and its interval crosses zero. Native history therefore provides one robust and one directional-but-inconclusive realization, not a two-realization replication. This valid endpoint is distinct from the failed paired-response contract above.

Fixed-update cross-world-companion readout. The four shared-answer-block fixed-update checkpoints answer the same 96 held-out two-question rows once under greedy no-thinking decoding with a 24-token cap. Ordered-pair exact correctness requires the exact two-line schema and both labels correct. For R1, cross-world and same-world score 85.42% and 68.75%, a $+16.67$-point contrast ($[+6.25, +27.08]$). For R2, they score 85.42% and 60.42%, a $+25.00$-point contrast ($[+15.63, +34.38]$). Schema validity is 96/96 in both first-realization arms and the second-realization cross-world arm, and 95/96 in the second-realization same-world arm; its remaining response terminates at the length cap.

Training-exposed and training-unseen same-world companions. Both controls reuse the four shared-answer-block checkpoint-250 artifacts and all 192 held-out targets, retaining each target in its original slot under the same two-line consumer, greedy no-thinking decoder, and 24-token cap. The training-exposed control selects the different-condition cell from the target's native theory with minimum hash under the frozen label-blind rule. Its cross-minus-same target-slot contrasts are -1.56 $[-5.73, +2.08]$ and -4.17 $[-8.33, 0.00]$. All 768 exposed-control requests return once; one R2 cross-world-trained response is schema-invalid and length-terminated and remains in the denominator.

The training-unseen control replaces every exposed companion with a unique same-world query absent from both SFT training arms and from the held-out target set; no companion is reused, and selection is label- and prediction-blind under a rule frozen before output inspection. Relative to the exposed condition, 50 replacements match the exact structural condition, 136 match polarity and arity, three match polarity, and three use the deterministic fallback; target slots remain balanced 96/96.

Table 6: **Consumer dependence persists with training-unseen companions.** Each accuracy uses all 192 targets. Accuracy rows are percentages; changes and paired 95% intervals are percentage points. Γ_s is a cross-consumer interaction, not a within-consumer arm contrast.

Quantity	R1	R2
Unseen same-world accuracy (%)		
cross-trained / same-trained	93.23 / 95.31	91.15 / 95.83
Cross-trained $U - F$	$+0.52$ $[-2.08, +3.13]$	-1.56 $[-3.65, 0.00]$
Same-trained $U - F$	$+12.50$ $[+7.81, +17.71]$	$+20.83$ $[+15.10, +27.08]$
Interaction Γ_s	$+11.98$ $[+6.25, +17.71]$	$+22.40$ $[+16.15, +29.17]$
Same-trained $U - E$	$+1.04$ $[-1.56, +4.17]$	$+0.52$ $[-1.04, +2.60]$

All 768 responses arrive in 768 attempts, satisfy the exact two-line schema, and stop normally, with no target filtering, companion reselection, parser repair, regeneration, or retry.

Let $F_{s,a}$, $E_{s,a}$, and $U_{s,a}$ denote target-slot accuracy under cross-world, training-exposed same-world, and training-unseen same-world companions for realization s and training arm a. The prespecified primary interaction is

$$\Gamma_s = (U_{s,\text{same}} - F_{s,\text{same}}) - (U_{s,\text{cross}} - F_{s,\text{cross}}), \tag{2}$$

judged against a $+5$-point threshold fixed in advance. Intervals use 10,000 paired target resamples within the same six source-block–by–target-slot strata, with frozen seed 3190914217; they quantify target-sampling uncertainty for fixed checkpoints and consumers.

Both interactions exceed the prespecified threshold with lower bounds above zero, and both same-trained $U - F$ changes are positive. Both same-trained $U - E$ intervals lie inside the prespecified $[-5, +5]$-point band for these fixed consumers. On the complete denominator, the same-trained checkpoints correct 25 targets and lose one relative to F in R1, and correct 42 and lose two in R2; the cross-trained counts are 4/3 and 0/3. Relative to E, the same-trained counts are 5/3 and 2/1. Exact companion reuse is therefore not necessary for this interaction. Because companion identity, provenance, and structural composition change together, the comparison does not isolate any one of these properties.

Paired organization on SFT-unseen theories. The companion controls above retain target theories represented in SFT through other conditions. This readout instead uses the unseen-theory population of Appendix B and the same four Qwen3-4B shared-answer-block checkpoint-250 artifacts. Target slots, instructions, the exact two-line response schema, the no-thinking Qwen3 template, temperature zero, and the 24-token cap remain fixed. The native organization keeps native companion assignments; the crossed organization swaps companions inside each of 96 complete two-theory clusters.

For realization s, training pairing t, and consumer organization o, let $A_{s,t,o}$ denote full-schema target-slot accuracy. In the organization subscript, same denotes native requests and cross denotes crossed requests. The prespecified interaction is

$$J_s^{\text{unseen}} = (A_{s,\text{same,same}} - A_{s,\text{same,cross}}) - (A_{s,\text{cross,same}} - A_{s,\text{cross,cross}}). \tag{3}$$

Intervals use 10,000 paired bootstrap resamples of the 96 complete mutual-swap clusters (seed 2533350915), retaining the aligned observations across training arms and organizations. The frozen transfer rule requires at least 95% schema validity in every cell, a maximum target-slot accuracy above 70% in each realization, both interactions to reach $+5$ points with lower confidence bounds above zero, and both same-trained organization changes to be positive.

All 1,536 planned responses are received once without content-dependent retry. Six checkpoint-by-organization partitions produce 192/192 exact two-line responses; two produce 191/192 because one response in each ends at the length limit. Every cell retains its full 192-target denominator, treating invalid responses as incorrect. The maximum target-slot accuracy is 96.88% in both realizations, and both satisfy every frozen transfer criterion.

Relative to crossed organization, the same-trained checkpoints correct 24 targets and lose none in R1, and correct 32 and lose one in R2. The cross-trained checkpoints correct one and lose five in R1, and correct two and lose none in R2. These transitions show the same-trained improvements underlying

Table 7: **Complete SFT-unseen paired-organization readout.** Accuracy entries give crossed/native request organization in percent. Within-checkpoint changes are native minus crossed; arm gaps are cross-world-trained minus same-world-trained under crossed/native requests. Both organizations contain the same 192 targets and companions. The interaction is $J = \mathrm{gap}_{\mathrm{cross}} - \mathrm{gap}_{\mathrm{native}}$. At 4B, J is the prespecified primary; at 8B, the crossed arm gap is primary and J secondary. Arm-gap intervals appear for the 8B primary; within-checkpoint changes and interactions have mutual-swap-cluster bootstrap 95% intervals.

		Accuracy, crossed/native (%)		Arm gap
Scale	Run	Cross-trained	Same-trained	Crossed/native
4B	R1	95.31 / 93.23	84.38 / 96.88	$+10.94$ / -3.65
4B	R2	93.75 / 94.79	80.73 / 96.88	$+13.02$ / -2.08
8B	R1	96.35 / 95.83	81.25 / 95.83	$+15.10$ $[+8.85, +21.35]$ / 0.00
8B	R2	95.83 / 96.35	81.77 / 94.79	$+14.06$ $[+8.85, +19.79]$ / $+1.56$

		Within-checkpoint change, native−crossed (pp)		Interaction (pp)
Scale	Run	Cross-trained	Same-trained	J
4B	R1	-2.08 $[-4.69, 0.00]$	$+12.50$ $[+8.32, +17.19]$	$+14.58$ $[+9.90, +19.79]$
4B	R2	$+1.04$ $[0.00, +2.60]$	$+16.15$ $[+10.94, +21.88]$	$+15.10$ $[+9.90, +20.83]$
8B	R1	-0.52 $[-3.12, +1.56]$	$+14.58$ $[+8.85, +20.83]$	$+15.10$ $[+8.85, +21.37]$
8B	R2	$+0.52$ $[-1.04, +2.60]$	$+13.02$ $[+7.81, +18.75]$	$+12.50$ $[+6.77, +18.75]$

the observed ranking reversal. This population has no singleton readout, and the organization change also changes whether a theory is repeated within the request; the result cannot separate those request properties.

The same readout at Qwen3-8B. The four R1-8B/R2-8B checkpoints answer the identical population under the identical two organizations, adding 1,536 one-attempt responses. All arrive without retry or missingness; 1,535 satisfy the exact two-line schema, the least valid partition is 191/192 (99.5%), and the maximum cell target-slot accuracy is 96.35%. The prespecified primary is the training-arm contrast under crossed requests: $+15.10$ points ($[+8.85, +21.35]$) and $+14.06$ points ($[+8.85, +19.79]$). Under native organization these gaps become a tie (184 versus 184 of 192) and a 1.56-point residual (185 versus 182 of 192), so the checkpoint order does not reverse. The prespecified secondary interactions are $+15.10$ points ($[+8.85, +21.37]$) and $+12.50$ points ($[+6.77, +18.75]$), using 10,000 paired mutual-swap-cluster resamples with seed 3471126501. Relative to crossed organization, the same-trained checkpoints gain 28 and 25 net correct targets while the cross-trained checkpoints change by -1 and $+1$. These checkpoints have no singleton readout, and the organization change again couples companion assignment with within-request theory repetition.

Table 7 preserves the complete fixed checkpoint readout underlying Figure 4.

E.2 ContractNLI context and evidence deletion

ContractNLI fixed-update construction. The contract study uses two independent Qwen3-4B fixed-update realizations of one contrast. Each arm contains the same 20,000 atomic contract–hypothesis judgments in 10,000 two-judgment rows and receives 250 optimizer updates at global batch 40; only checkpoint 250 is read. In each persisted row, `input` is empty, one user instruction contains `[1]` contract plus target hypothesis h followed by `[2]` contract plus companion hypothesis g, and one assistant response emits `1.` `<label>` then `2.` `<label>` on separate lines. The independent realization reuses the same two immutable training corpora, so both use this shared indexed-answer-block serialization. Within every update, the two arms match the ordered first-position sequence, second-position label sequence, complete occurrence multiset, and response strings. The change is whether the two co-sequenced judgments come from the same contract or different contracts. The original eligibility rule deliberately selects contracts A and B with different official labels for target hypothesis h and the same label for distinct companion hypothesis g, then emits reciprocal same- and cross-contract rows. Every cross-contract row therefore has 0% target-hypothesis agreement by design, while same-contract rows agree by construction; the companion's

label on h is not one of that row's two explicit answers, although reciprocal rows can supervise that judgment elsewhere in the corpus. The repaired-relation test below evaluates the behavioral prediction arising from this construction-level relation; it does not identify a learned or neural mechanism. The independent pair uses a different model seed and a prospectively changed update-block order; both checkpoints pass physical loadability checks and are read only with the official-test constructions already frozen for the original pair. The deletion requests are later issued to that pair byte-for-byte under their separately frozen rule.

ContractNLI paired-use readout. On development, we construct native, crossed, and no-document requests for 338 target judgments in each condition across 24 document-pair clusters. Each checkpoint answers 1,014 requests once, for 2,028 responses across the pair. The reported interaction uses only native and crossed organization; the no-document cell is diagnostic. The development interaction is $+13.91$ points ($[+10.51, +17.81]$). Its interval uses 10,000 paired resamples of the 24 clusters (seed 3347241901). Before the official test contracts were opened, we recorded the same deterministic construction and one-attempt readout for 732 targets in each condition across 51 document-pair clusters. The test interaction is $+12.98$ points ($[+9.54, +16.57]$). The test interval uses 10,000 document-pair cluster resamples with seed 4242202610 while retaining the sampled target denominator. The no-document test cells score 46.45% and 44.81%, well below the native and crossed cells, so the paired-use interaction is not a restatement of document-independent base competence. The independent realization answers the same 732 targets in each condition across the same 51 document-pair clusters. Its cross-contract-trained checkpoint moves from 73.50% native to 87.57% crossed, and its same-contract-trained checkpoint from 84.56% to 83.61%, an interaction of $+15.03$ points ($[+11.48, +18.87]$). Cluster contributions are positive in 20 of 24 development clusters for the original pair with four zero, in 42 of its 51 test clusters with six zero, and in 43 of 51 independent- pair test clusters with five zero.

ContractNLI source-agreement readout. The development readout presents 356 units per condition across 20 deterministic three-contract units, or 712 requests per checkpoint and 1,424 across the pair. Each trio contributes equally to the reported cells and interaction; the development interaction is $+22.30$ points ($[+17.29, +27.25]$). The interval uses 10,000 paired trio resamples (seed 4141202609). One strict-schema-invalid response remains in the full denominator and is counted wrong. The frozen test readout presents 783 units per condition across 38 document trios, again with equal-trio weighting and 10,000 paired trio resamples (seed 4242202611). Agreeing and conflicting requests use different companion contracts, so contract identity, wording, length, evidence location, and difficulty vary together with label agreement. For the original pair, each checkpoint answers 3,762 requests across the two test readouts; all 7,524 requests return, 7,513 outputs satisfy the strict schema, and invalid outputs remain in the denominator and count wrong. The independent realization answers the same 783 units per condition across the same 38 document trios: 90.04% conflicting and 69.81% agreeing for the cross-contract-trained checkpoint, and 86.78% and 87.73% for the same-contract-trained one. The interaction is $+21.18$ points ($[+17.03, +24.57]$). For the original pair, 19 of 20 development trio contributions and 37 of 38 test contributions are positive, with one zero in each split; for the independent pair, 37 are positive and one is negative. The original pair follows a protocol recorded before the test split was opened; the independent pair changes the training realization.

ContractNLI companion-evidence deletion. The deletion returns to the original checkpoint pair and edits only the companion contract of an agreeing request. One condition removes the companion evidence consistent with the target label; the other removes a length-matched span irrelevant to the target judgment. The target contract, target hypothesis, companion hypothesis, response format, decoder, and scoring are unchanged, and the rule is fixed in advance. Because the official test split had already been opened by the two unedited readouts, this continues validation on that split under frozen rules rather than supplying a first look at it. Cells weight each three-contract unit equally. The development and test interactions are -37.71 points ($[-59.58, -16.88]$) and -23.49 ($[-38.55, -12.84]$); on test, 17 of 33 trio contributions are negative and 16 are zero.

Aligning the edited outputs with the undeleted agreeing requests over the same 33 test trios gives cross-contract-trained accuracies of 75.97%, 73.70%, and 93.56% for original, neutral deletion, and target-evidence deletion. Of the 18 units where target-evidence deletion beats the neutral deletion, 16 were already answered incorrectly on the undeleted request and two were introduced by the neutral deletion. Companion answers are unchanged throughout on development; on test they are unchanged

for the same-contract-trained checkpoint and for 92 of 96 cross-contract-trained requests. One frozen official-test deletion unit, `e1b7e7db36a5eed04187d468` retains contract 540's target return clause and removes only the parallel clause in companion contract 77: the cross- contract-trained target changes from `NotMentioned` to `Entailment`, its companion answer remains `NotMentioned`, and the same-contract-trained checkpoint was already correct. The target excerpt below preserves both the notice trigger and archival exception in every request:

> Promptly upon notice from either party that it does not wish to proceed with the Transaction, each party shall return to the other party or destroy all copies of the Evaluation Material [...] except that outside counsel to the receiving party may retain one copy of the Evaluation Material for archival purposes and solely to comply with applicable law, rule or regulation.

The neutral deletion removes the complete passage:

> The parties will entrust confidential information solely to those employees, consultants and third party companies which are concerned with the processing of the order and which are legally or contractually obligated to secrecy.

The target-evidence deletion instead removes this complete return provision:

> 9. At the end of the contract the receiving party shall return the confidential information and any possibly extant copies after receipt of a written request, provided that this is technically feasible and does not violate the retention obligations.

Both edits retain the following sentence:

> As an alternative to that, the receiving party can also ensure in writing the destruction of the received documentation if that is technically feasible.

This is an auditable case of the aggregate dependence under a test-time edit, not a persisted training row, mechanism proof, or repair procedure.

The frozen deletion requests are then issued byte-for-byte to the independent pair: 96 agreeing official-test units in 33 trios, two deletion conditions, and 384 outputs, all strict-schema valid, with no retraining, re-editing, or unit reselection. For its cross-contract-trained checkpoint, original, neutral-deletion, and target-evidence-deletion accuracies are 75.82%, 75.82%, and 92.19%; for its same-contract-trained checkpoint they are 92.97%, 89.94%, and 89.94%. The deletion interaction is -16.37 points ($[-25.17, -9.01]$), and trio contributions are negative, zero, and positive in 14, 19, and 0 trios. All 96 neutral-deletion target predictions of the independent cross-contract-trained checkpoint equal its undeleted predictions, while target-evidence deletion produces 15 corrections and no regressions, all on targets already wrong before deletion. The same-trained target predictions are identical across the two deletion conditions, including one target lost relative to the undeleted request. Companion predictions change on one cross-trained and two same-trained requests. Six corrected units overlap the original pair's 18, leaving 12 original-only and nine independent-only corrections. These audits support transfer of the signed aggregate response, not a shared item-level error signature or a cost-free editing procedure.

E.3 Repaired-relation training and its costs

Repaired-relation training test. For each training realization, one repaired cross-contract schedule re-pairs the companion atoms while preserving every target occurrence, explicit two-question judgment, row position, optimizer update, per-update companion- atom multiset, token total, and the 250-update recipe. Target-hypothesis agreement changes from 0% to 42.75% and 42.67%. These schedules weaken rather than eliminate the relation and do not constitute independence or complete semantic decorrelation; companion identity, wording, length, and local layout change with the reassignment. The valid repairs change the physical companion document in 5,451 and 5,511 rows, respectively, distinct from the later constrained-sham extrema.

One new training run per schedule, together with the existing official-test paired-use, source-agreement, and deletion matrices, was frozen before repaired outputs were read. Both exact

checkpoint-250 artifacts qualified. Per repaired checkpoint, the frozen matrix contains 2,196 paired-use, 1,566 source- agreement, and 192 deletion requests; all 7,908 one-attempt responses across the two checkpoints return. One strict-schema-invalid output in the first realization remains in the denominator and counts wrong, with no retry or refill. The repaired-training claim reported here is limited to the prespecified source-agreement and deletion endpoints.

The equal-trio agreeing-minus-conflicting gaps change from -19.39 to -0.18 and from -20.23 to -0.33 points. The repaired-minus-original shifts are $+19.21$ ($[+16.14, +22.44]$) and $+19.90$ ($[+17.03, +22.65]$), positive in 37 and 36 of 38 trios with no negative trio. Agreeing-condition changes are $+15.79$ ($[+11.72, +19.50]$) and $+18.43$ ($[+15.75, +21.09]$), with wrong-to-correct versus correct-to-wrong counts of 146/17 and 156/12. Conflicting-condition changes are -3.41 ($[-5.82, -1.20]$) and -1.47 ($[-3.70, +1.24]$), with transition counts of 21/41 and 18/34; only the first decline is interval-supported. Across 783 paired targets, 779 and 778 predictions are invariant between agreeing and conflicting requests.

Repaired agreeing/conflicting equal-trio accuracies are $87.92\%/88.09\%$ and $88.24\%/88.57\%$ in the two realizations. On a descriptive equal-weight mix of those two conditions, the repaired checkpoints gain $+6.19$ and $+8.48$ points over their matched original cross-contract checkpoints; this mix was not a prespecified deployment utility. The prespecified useful-content guardrail was not met: it required agreeing-condition gains of at least five points in both realizations and conflicting-condition change intervals with lower bounds no worse than -3 points. The corresponding deletion criterion required at least five-point neutral-deletion gains and target-evidence- deletion change intervals with lower bounds no worse than -3 points. Failure of these criteria does not by itself establish an accuracy loss; the signed changes and their uncertainty distinguish observed costs from unestablished utility.

The target-evidence-minus-neutral deletion contrast becomes exactly zero in both repaired checkpoints. Relative to the matched originals, the shifts are -19.86 ($[-28.50, -12.07]$) and -16.37 ($[-24.85, -8.95]$). Neutral-deleted accuracy changes by $+16.39$ ($[+7.25, +26.02]$) and $+18.69$ ($[+10.26, +28.54]$), whereas target-evidence- deleted accuracy changes by -3.46 ($[-12.88, +5.91]$) and $+2.32$ ($[-4.70, +10.81]$). Each repaired checkpoint gives identical target predictions across the two deletion conditions on all 96 units.

Repaired-checkpoint paired-use competence. The frozen paired-use matrix contains the same 732 official-test targets under native, crossed, and no-document requests for each repaired checkpoint. All six cells return $732/732$ strict-schema-valid responses. Target accuracy is 85.66% (627/732), 85.38% (625/732), and 44.81% (328/732) in the first realization, and 84.70% (620/732), 85.11% (623/732), and 44.81% (328/732) in the second, respectively. Companion-slot accuracy under native and crossed requests is $93.58\%/95.77\%$ in the first realization and $92.90\%/96.17\%$ in the second, versus 45.36% without documents. Thus the repaired checkpoints retain document-conditioned paired- task competence while the source-agreement and deletion contrasts collapse. Removing both documents removes evidence for both requested judgments, so this matrix does not establish itemwise use of the non-target companion.

Common-support control feasibility. We prospectively tested whether a semantic repair and a 0%-agreement sham could change the same target-row support in every optimizer update and every exact companion-hypothesis/label bucket. The sham must preserve the original companion's exact three-way wrong truth on the target hypothesis; the repair may retain that truth or replace it with the supervised target truth. A change means a different physical companion document, not a different occurrence of the same document. The frozen qualification required at least 3,000 such changes and 2,500 repair agreements (a 25-point relation separation). Both schedules meet the support floor, but their maximum relation separations are only 21.37 and 20.13 points, below the frozen 25-point threshold. The study therefore stopped before model calls or training. This corpus cannot supply the specifically qualified matched sham, so the repaired test supports a construction-level relation explanation without identifying agreement as its unique cause.

Two complete-output units illustrate the movement and its cost; they are post-hoc cases, selected neither for prevalence nor as mechanism evidence. Unit `0365806522f75dad1f776b1e` asks whether the receiving party may independently develop similar information, which the target contract explicitly exempts. The original first-realization cross-contract checkpoint answers `NotMentioned` with an agreeing companion and `Entailment` with a conflicting one; both repaired checkpoints

Table 8: **Three natural-text signatures under matched re-pairing.** Agreement is the fraction of training rows in which the companion contract agrees with the target contract on the target hypothesis. Source gap is agreeing minus conflicting target accuracy; deletion contrast is target- evidence minus neutral deletion accuracy. These two repair endpoints and their five-point shift predictions were frozen before repair training. The question-cross interaction was specified later, after repair training and its main readout but before new requests on previously opened official-test documents. Shifts compare each repaired checkpoint with its matched original cross-contract checkpoint; brackets are paired 95% intervals. All effects other than training agreement are percentage points.

Measure	First realization	Second realization
Training agreement, original → repaired	$0.00\% \to 42.75\%$	$0.00\% \to 42.67\%$
Predicted before repair training		
Source gap, original → repaired	$-19.39 \to -0.18$	$-20.23 \to -0.33$
Matched gap shift [95% interval]	$+\mathbf{19.21}$ $[+16.14, +22.44]$	$+\mathbf{19.90}$ $[+17.03, +22.65]$
Deletion contrast, original → repaired	$+19.86 \to 0.00$	$+16.37 \to 0.00$
Matched contrast shift [95% interval]	$-\mathbf{19.86}$ $[-28.50, -12.07]$	$-\mathbf{16.37}$ $[-24.85, -8.95]$
Later frozen new-request check		
Question-cross interaction, original → repaired	$-17.32 \to 0.00$	$-22.25 \to 0.00$
Matched interaction shift [95% interval]	$+\mathbf{17.32}$ $[+10.89, +23.75]$	$+\mathbf{22.25}$ $[+17.42, +27.25]$

answer `Entailment` under both requests. Unit `0003f06f869adc4adff4f9d8` asks whether ordinary confidential information may be shared with representatives. The original first-realization checkpoint is correct in both conditions. The matched original second-realization checkpoint is already wrong under the agreeing companion but correct under the conflicting one. Both repaired checkpoints answer `Contradiction` in both conditions, so relative to the second realization the newly introduced error is confined to the conflicting condition.

E.4 Question-cross test with fixed physical companions

A post-hoc diagnostic of the earlier source-agreement outputs matched target document, gold label, companion question and the same physical neighbor pair across target questions. It motivated the separate frozen test below; that exploration is not a prospective confirmation, and varying the target question also varies its wording and evidence target.

Frozen crossed-question requests with fixed physical companions. After repair training and its primary readout had finished, but before the requests below, we froze a result-blind SHA-256 ordering of official-test documents, one grouping rule, an estimand, uncertainty procedure, and a joint success criterion. Disjoint document triples supply a target contract and two fixed physical companions C_1, C_2. An eligible group fixes the target contract and gold label, C_1, C_2, and a companion question with the same official label in both companions. Target questions qualify when C_1 and C_2 disagree on the target hypothesis and exactly one agrees with the target's gold answer; both C_1-agree and C_2-agree directions must occur. The target question and the evidence it asks about necessarily change across these directions. No prediction or measured effect enters selection. The carrier contains 38 disjoint trios, 99 groups, 80 target documents, and 406 target-question units. Each of four already-trained checkpoints receives both companion requests for every unit: 812 requests per checkpoint, with 3,248/3,248 one-attempt strict-schema-valid outputs and no retry or refill. The documents and checkpoints had already been opened in earlier tests; the new request combinations, not a document split, are prospective.

For a target question i, let $d_i = 1[\text{correct}(C_1)] - 1[\text{correct}(C_2)]$. Within group g, define T_g as one-half the mean d_i for C_1-agree questions minus one-half the mean d_i for C_2-agree questions. Average T_g equally within each trio, then average the 38 trios equally to obtain T. Percentile 95% intervals use 10,000 paired whole-trio bootstrap resamples across all four checkpoints. The frozen joint criterion required each original cross-trained $T \leq -5$ points with upper interval bound below zero and each matched repaired-minus- original shift $\geq +5$ points with lower bound above zero. Both original values and matched shifts satisfy it (Table 8); the original T intervals are $[-23.75, -10.89]$ and $[-27.25, -17.42]$ points, and the repaired intervals are exactly $[0.00, 0.00]$ in each realization.

Table 9: **Original-schedule ContractNLI accuracy levels behind the behavioral results.** Target accuracy is in percent. Repaired schedules are reported through their prespecified contrasts and shifts in Table 8 and the paired-use competence protocol above. Paired organization uses 338 development targets in 24 document-pair clusters and 732 test targets in 51 clusters. Source agreement uses 356 development units in 20 document trios and 783 test units in 38 trios, with equal weight per trio. Companion deletion uses 96 agreeing test units in 33 trios; the same units and deletion requests are issued byte-identically to both training realizations. Independent denotes the second Qwen3-4B training realization, read only on test with frozen constructions.

Paired organization			
Split, pairing	Checkpoint	Native	Crossed
Dev, original	Cross-contract	76.04	85.80
Dev, original	Same-contract	85.50	81.36
Test, original	Cross-contract	77.32	87.02
Test, original	Same-contract	85.93	82.65
Test, independent	Cross-contract	73.50	87.57
Test, independent	Same-contract	84.56	83.61
Source agreement			
Split, pairing	Checkpoint	Conflicting	Agreeing
Dev, original	Cross-contract	85.83	71.01
Dev, original	Same-contract	82.23	89.70
Test, original	Cross-contract	91.51	72.12
Test, original	Same-contract	84.66	88.81
Test, independent	Cross-contract	90.04	69.81
Test, independent	Same-contract	86.78	87.73
Companion deletion			
Split, pairing	Checkpoint	Neutral	Target-evidence
Dev, original	Cross-contract	52.71	89.17
Dev, original	Same-contract	88.33	87.08
Test, original	Cross-contract	73.70	93.56
Test, original	Same-contract	93.30	89.66
Test, independent	Cross-contract	75.82	92.19
Test, independent	Same-contract	89.94	89.94

The latter are not rounded near-zero effects: original target answers differ between C_1 and C_2 for 92/406 and 98/406 units, whereas repaired target answers are identical for 406/406 units in both realizations. Repaired companion answers still differ for 34/406 and 30/406 units; target invariance is not whole-output invariance or proof of correct evidence routing.

The negative crossed interaction rules out a target-question-independent fixed companion offset on the defined correctness-difference scale, because that offset cancels between the two agreement directions. It does not isolate abstract agreement: question wording and its evidence target also vary. The repair changes 74 and 91 of the 812 target response cells from wrong to correct, while changing 26 and 29 previously correct cells to wrong. A zero interaction here is therefore not uniformly improved accuracy.

E.5 RACE RESPONSE FORMAT AND PAIRED-USE SCOPE

RACE fixed-update construction and readout. RACE (Lai et al., 2017) uses passage-level same-world pairing and answer-letter-bucket cross-passage rewiring under the shared-answer-block layout. Each arm contains 10,000 rows and 20,000 unique occurrences with seed/data seed 1662733758, global batch 40, 250 updates, and checkpoint 250. Within every update, the two schedules match the ordered first-position sequence, second-position occurrence multiset, complete occurrence multiset, response strings, and trained-token histogram; only the second-position partner is reassigned across passages within answer-letter class. The complete dev readout has 4,887 questions in 1,389 passage clusters. Each checkpoint answers once with greedy no-thinking decoding under the same single-letter request and eight-token cap; raw generations are retained. Intervals use 10,000 paired passage-cluster bootstrap resamples (seed 338046104).

Table 10: **Descriptive decomposition of the Qwen3-8B RACE contrasts.** The phenotype is defined by a post-training same-world output, so the columns are accounting terms rather than a causal mediation; "other rows" is an arithmetic remainder, not a form-independent residual effect.

Endpoint	Full contrast	Phenotype rows	Other rows	Share
Exact and correct letter	$+28.63$ (1,399)	$+26.07$ (1,274)	$+2.56$ (125)	91.07%
First standalone letter	$+28.93$ (1,414)	$+26.40$ (1,290)	$+2.54$ (124)	91.23%
Exact one-letter compliance	$+30.24$ (1,478)	$+30.02$ (1,467)	$+0.23$ (11)	99.26%

For Qwen3-8B, cross minus same is $+28.63$ $[+27.12, +30.09]$ for exact-and-correct accuracy, $+28.93$ $[+27.44, +30.42]$ for first standalone letter correctness, and $+30.24$ $[+28.77, +31.72]$ for exact one-letter compliance. The Llama-3.1-8B-Instruct realization reuses the same frozen arms and schedule. Every response in both Llama arms continues the trained two-line response format with multiple standalone letters and terminates at the token cap, so exact-and-correct accuracy and exact one-letter compliance are both 0.00% in both arms. First-letter accuracy is 79.91% cross and 26.64% same, a $+53.26$-point contrast $[+51.67, +54.84]$. Because neither arm accepts the shared strict contract, this realization establishes no arm ranking on that endpoint; the first-letter diagnostic cannot substitute for it.

Raw-output audit of the Qwen3-8B RACE readout. Across the complete 4,887-question population, the cross-world and same-world checkpoints respectively return 4,392 versus 2,993 exact-and-correct, 4,412 versus 2,998 first-standalone-letter correct, and 4,861 versus 3,383 exact one-letter-compliant responses.

In the same-world arm, 1,485 responses (30.39%) consist of exactly eight line-feed characters. Every one is marked `finish_reason=length`, uses all eight completion tokens and contains no standalone option letter. The cross-world arm contains no response of this form. More broadly, 3,398 same-world responses (69.53%) reach the cap, against 17 (0.35%) cross-world responses. Table 10 decomposes the frozen contrasts over the questions selected by the same-world line-feed phenotype.

Native paired-answer readout of the same Qwen checkpoints. A separate result-blind readout presents two complete RACE problems before one assistant response. Every one of the 4,887 development questions appears once in target slot 1 and once as an answer-independent companion in slot 2; the next question in the same passage is used when available, and the 31 singleton passages form a fixed cycle. Both checkpoint-250 artifacts receive byte-identical requests with the training renderer, greedy no-thinking decoding, and a 24-token cap. Every request is answered once and retained in the denominator.

For this fixed checkpoint pair, cross-world and same-world target-slot accuracies are 90.67% and 90.06%; their $+0.61$-point difference has a 10,000-replicate target-passage-cluster bootstrap interval of $[-0.08, +1.33]$ (seed 1970400636). Companion accuracy is 91.14% and 90.67%, both-slots-correct is 83.20% and 82.18%, and exact two-line format is 99.94% and 99.98%. Three cross-world responses and no same-world responses terminate at the cap; neither arm emits an empty or pure-newline response. The primary interval lies wholly inside the frozen ± 2-point band while both format and length floors pass, so this is a valid bounded null for the fixed artifacts and consumer, not evidence of general equivalence.

Same- versus cross-passage use factorial. We separately test whether the preceding fixed-pair contrast depends on sharing the target passage at readout. On the official RACE test split, we symmetrically exclude 42 questions whose passages have no second question, leaving the same 4,892 targets in 1,365 passage clusters for all four cells. The same-passage condition uses a deterministic next-question cycle. The cross-passage condition bijectively reassigns the identical companion multiset to another passage while matching answer letter and RACE level. Target identity, order, gold labels, request template, decoder, parser, and 24-token budget are fixed across use conditions.

Under cross-passage use, target-slot accuracy is 90.04% for cross-world training and 89.35% for same-world training, a $+0.70$-point simple contrast ($[-0.04, +1.42]$). Under same-passage use, the corresponding values are 90.17% and 89.49%, a $+0.67$-point simple contrast ($[-0.06, +1.38]$). These simple contrasts are descriptive. The prespecified decision-bearing estimand is their difference:

$+0.020$ points. Because a result-blind audit found extensive reuse of cross-companion passages, its 95% interval uses an intercept-only repeated-dyad cluster-robust sandwich with target and companion passages as members; the interval is $[-0.349, +0.390]$ points. A target-passage bootstrap sensitivity gives $[-0.349, +0.388]$. All 19,568 responses are received, every cell exceeds 99.95% exact two-line format, and length termination occurs in at most one row per cell. The primary interval lies wholly inside the frozen ± 2-point band, yielding a bounded null for this training-by-request-organization interaction—not general equivalence or evidence that all resource and consumer interactions are absent.