# SAME KNOWLEDGE, DIFFERENT USE: HOW TRAINING PAIRINGS SHAPE COMPOSITIONAL GENERALIZATION IN LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Fine-tuning can teach a language model to recall a fact without teaching it to apply that fact in a new problem. We study how the assignment of training examples contributes to this gap. Models first learn a recorded symbol for each of 384 fictional entities and learn to classify six-symbol sequences when all inputs are supplied. An application question supplies only the first five symbols and names the entity whose record specifies the sixth. The target is a category label denoting a range of the final-state probability under a fixed two-state hidden Markov model. We compare two training conditions using identical starting weights within each comparison. Each of 96 application-trained names receives eight different five-symbol contexts. In one condition, all eight contexts require the same answer for that name: SAME8 (8:0). In the other, four require one answer and four another: CROSS8 (4:4). We construct this difference by exchanging contexts between names with the same recorded symbol. The overall application material, labels, number of contexts per name, factual and explicit-task replay, and training updates remain matched. On contexts withheld from the compared application training, 4:4 improves accuracy by an average of 36.8 percentage points across three Qwen worlds, 38.4 points in Llama, and 39.8 points across two Phi assignments sharing one starting checkpoint and world. These gains do not accompany better factual recall. A subsequent Qwen comparison shows that 7:1 training already produces 29–36-point gains over 8:0. Its single minority-answer context is presented six times, and equivalence to 4:4 is not established. The original 4:4 advantage persists when both conditions' application labels uniquely identify the recorded symbol. Transfer to names without application examples is less consistent, and better ordinary application can coexist with worse performance when instructions require using a conflicting supplied symbol. Replacing selected feed-forward components between the original trained copies partially improves answers without identifying a unique composition algorithm. The results show that how the same available application material is assigned can substantially change generalization.

## 1 INTRODUCTION

A language model may recall a fact and solve a problem when that fact is supplied, yet fail when it must use the learned fact with new inputs. Prior work documents related gaps in compositional questions and controlled knowledge manipulation (Press et al., 2023; Allen-Zhu & Li, 2025; Dai et al., 2026). We ask how training contributes to this gap: can assigning the same available application material differently change whether a model uses its records successfully?

Our comparison separates different questions from different answers. Each application-trained name receives eight distinct input sequences. In one condition, all eight require the same answer for that name; we call this **SAME8 (8:0)**. In the other, four require one answer and four another; this is **CROSS8 (4:4)**. A constant answer associated with the name can fit the first condition's training questions despite their different inputs. We create the second condition by reassigning sequences between names with the same relevant fact, so every target remains correct.



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.

The compared models start from identical weights within each comparison and receive matched overall application material, factual examples, fully supplied task examples and training updates. This extends work on training organization (Lampinen et al., 2025) with a controlled change in which names accompany which inputs. The 4:4 assignment improves held-out application across the tested Qwen, Llama and Phi comparisons. In three Qwen worlds, a subsequent 7:1 assignment already gives a large benefit. Its one distinct minority-answer context is presented six times; equivalence to 4:4 is not established.

We then distinguish what improved: answers for names with equivalent records, transfer to names without application examples, previously unpaired name–answer combinations, and compliance with conflicting supplied inputs. Replacing internal model components partially changes the original learned behavior. The contribution is the controlled assignment experiment and these measured consequences. "Same knowledge" means matched factual supervision and separately measured recall, not identical internal representations. Application labels can themselves reveal or reinforce a fact.

## 2 TASK AND TRAINING COMPARISON

### 2.1 ONE LEARNED INPUT IS OMITTED FROM THE QUESTION

Consider the fictional entity `Analy Lazy`, whose recorded final observation is $Q$. An application question supplies five symbols, $S, Q, Q, P, Q$, and refers to `ARCHIVE(Analy Lazy, 2020)` for the sixth. The completed sequence is therefore $S\ Q\ Q\ P\ Q\ Q$, whose reference answer is category $A$. We separately ask whether the model can recall $Q$, classify the sequence when $Q$ is supplied, and classify it when the question only references the record.

These are reference answers, not claimed responses of a model. The archive reference is text in the question; it executes no external lookup. Only the sixth observation comes from the taught record. The first five are hypothetical inputs, not the entity's earlier history. The fixed date identifies the record, and no competing dated updates are tested. Appendix A gives the exact prompts.

A **factual world** is a generated collection of records. It contains 384 fictional entities in 192 pairs whose members share a recorded six-symbol history. A **context** is the five-symbol sequence supplied in an application question. The missing sixth observation is the **recorded symbol**; its inherited implementation name is *endpoint*. Record symbols $P, Q, R, S$ differ from answer categories $A, B, C, D$.

A fixed two-state hidden Markov model (HMM) defines the target. Its unobserved state evolves while emitting observation symbols. Exact filtering of six observations gives the probability that the final state is 1. The labels $A, B, C, D$ denote the intervals $[0, .25)$, $[.25, .5)$, $[.5, .75)$ and $[.75, 1]$. For the example above, that probability is approximately $0.0633513$, giving label $A$. The language model predicts a category, not this probability or a reasoning trace. Its token probabilities are distinct from the HMM probability. Figure 1 separates model prediction from reference scoring; Appendix A retains the matrices and filtering equations.

### 2.2 CHANGE THE ASSIGNMENT WHILE MATCHING THE MATERIAL

Before the comparison, models learn records and classification with all six inputs supplied, with continued factual replay. The resulting **starting checkpoint K** is a saved model. Compared copies independently continue from identical K weights within a comparison. Earlier preparation differs across model families; this is not a comparison of identically acquired models or an architecture ranking (Appendix B).

For two names sharing a record, choose eight contexts requiring answer $a$ and eight requiring $b \neq a$. SAME8 gives the first eight to one name and the second eight to the other. CROSS8 exchanges the names on four paired positions. Each name still receives eight distinct contexts, now requiring four answers of each kind. The context–recorded-symbol–target combinations match when the name is disregarded. The complete name–context examples intentionally differ.

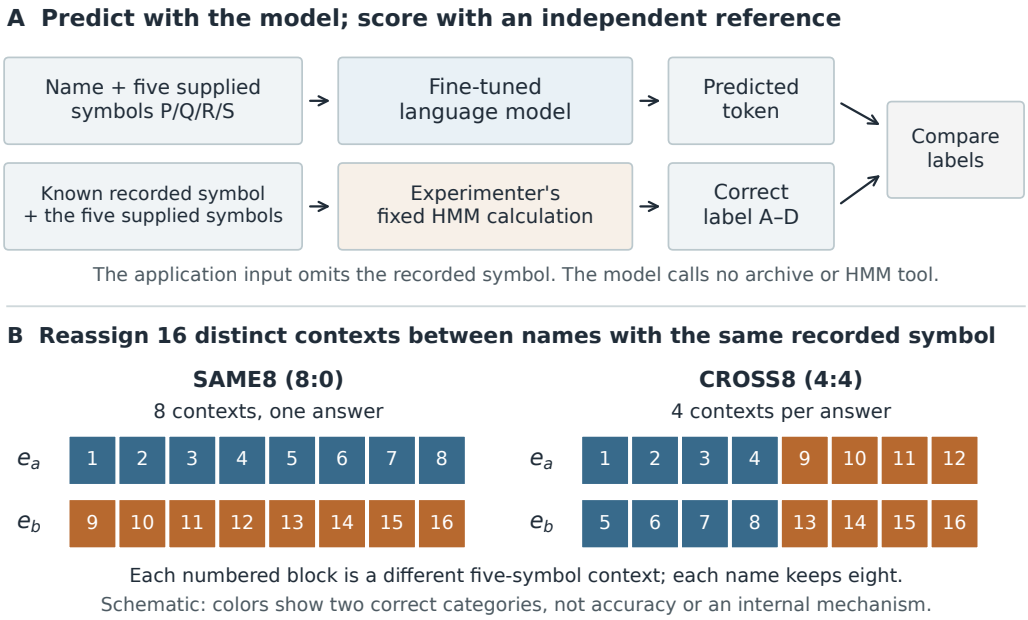



Figure 1: Task and matched assignment (schematic). Model prediction and reference scoring are separate; the model receives no recorded symbol and calls no archive or HMM tool. Predictions use the full vocabulary, with invalid outputs counted as errors. Each name keeps eight distinct contexts. SAME8 (8:0) assigns one answer per name; CROSS8 (4:4) assigns four contexts per answer. Colors denote reference categories, not measured accuracies. Reassignment preserves context–recorded-symbol–target combinations.

Application training covers 96 names, while factual replay covers all 384. Each copy receives 18,432 examples and 2,304 updates, with matched factual and fully supplied classification material. Repeated prompt wordings give 48 application presentations per trained name without adding distinct contexts. Appendix B preserves the optimizer, loss, seed and token details. Reassignment changes several associations and may reveal the recorded symbol through labels; answer entropy is not isolated as the sole cause.

**Intermediate assignments.** We subsequently tested 7:1 and 6:2 in the same three Qwen worlds. Fixed, nested choices exchange one or two of the four positions exchanged by CROSS8, independently of new model outcomes. Each new model starts from its world's original K, not another application-trained copy. Historical 8:0 and 4:4 models are not retrained. Eight contexts and 48 presentations per name remain fixed: 7:1 has six minority-context presentations; 6:2 has twelve. Training dose, auxiliaries and within-world seed are unchanged, with one trained model per new condition/world. This extends previously inspected worlds and test support, not a fresh independent replication (Appendix H).

## 3 What the Evaluations Measure

We evaluate Qwen3-4B in worlds W3, W1 and W4, Llama-3.2-3B-Instruct in W3, and Phi-4-mini-instruct in two assignments sharing one acquired K and W3 records (Yang et al., 2025a; Meta, 2024; Abouelenin et al., 2025). Phi changes assignment and optimization seed, not factual world. Families and repeated assignments do not add independent Qwen worlds.

**Three capabilities and distinct populations.** **Recall** asks for the recorded symbol alone. **Fully supplied classification (EXPLICIT)** provides all six symbols. **Omitted-input application (DIRECT)** supplies five symbols and the name reference. These test reporting a fact, solving the task with it visible, and answering with it omitted. They do not identify the model's internal sequence of operations.

Ordinary tests use eight contexts per name. We separately evaluate the 96 application-trained names and 192 **fact-only test names**, whose records were taught without application examples. A further 48 fact-only names form an additional test population, and 48 are reserved for validation. The legacy label "unexposed" does not mean absence of factual learning. Test contexts were excluded from the compared application curricula, although some strings appeared in earlier fully supplied training. Researcher inspection and renderer exposure are recorded separately.

Correctness uses the highest-scoring next token over the full vocabulary. Recall requires the correct symbol in $P, Q, R, S$; classification requires the correct category in $A, B, C, D$. Invalid outputs count as errors. We compare trained copies with each other and with K: exceeding 8:0 need not mean exceeding the starting model. The supplementary recall-then-classify route (SELF), including its selection and wrong-recall handling, remains in Appendix B.

**Tests of what the answers depend on.** For two names with the same record and a fixed context, the correct answer is identical. We report how often both answers are correct and their **total variation**, a distributional difference over the four answer labels plus the probability mass on all other vocabulary tokens. Smaller variation alone is insufficient: both answers could agree and be wrong.

A separate balanced grid asks whether a model could succeed by ignoring the name's record. It combines 32 contexts with two names for each recorded symbol, giving the four symbols equal weight. If $F(c, v)$ is the reference category for context $c$ and recorded symbol $v$, the best context-only accuracy is

$$B_0 = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \max_{y} \frac{1}{4} \sum_{v \in \{P,Q,R,S\}} \mathbf{1}\{F(c, v) = y\}. \tag{1}$$

This selects the most frequent reference answer for each context, so the bound need not be 25%. It applies only to that grid. Exceeding it excludes a strictly context-only predictor there, not a particular alternative retrieval algorithm. The **conflicting-input test** instead supplies an alternative sixth symbol that changes the target and instructs the model to use it. That supplied value is the intended input, not an error in the question.

**Uncertainty and numerical qualification.** Application questions share recorded histories and visible contexts. History-cluster intervals resample original history pairs; a crossed sensitivity also reweights recurring contexts. Intervals condition on the fitted checkpoints and observed support, not independent training repetitions. Qwen's summary weights its three worlds equally; Phi's two-assignment mean is descriptive. Appendices C, F and H.3 preserve the different resampling rules. The graded study additionally adjusts nine adjacent comparisons, three per world; secondary comparisons are unadjusted.

The extension failed its original BF16 checks for agreement across batch compositions and with native generation. BF16 is a reduced-precision number format. After these failures, the amended policy retained recorded batched estimates and rescored all fifteen Qwen conditions **one question at a time** (batch size one). Both paths use identical checkpoints and the matched evaluation cases used in this extension. The principal 7:1 advantage survives both paths, while some secondary classifications differ. This sensitivity does not restore the original failed gates (Appendix H.2).

## 4 RESULTS

**Application improves without better recall.** On new contexts for application-trained names, CROSS8 exceeds SAME8. Qwen's average gain is 36.8 percentage points across three worlds. Llama gains 38.4 points; the two Phi assignments have a descriptive mean of 39.8 points. Every original comparison also improves above K (Figure 2). Exact effects and both interval schemes remain in Table 4.

In Phi's first assignment, both copies recall all 384 measured facts. Fully supplied classification is 97.14% under SAME8 and 95.18% under CROSS8, yet omitted-input accuracy rises from 41.67% to 80.21% (Table 1). The gain therefore does not accompany better recall or fully supplied performance. Other capabilities are not uniformly preserved: Qwen W1 CROSS8 has five recall errors, and Llama's fully supplied validation accuracy falls from 96.48% to 91.28%. Table 6 retains both validation and test results with uncertainty.

Table 1: Record recall and application accuracy (%). Each application-trained name has eight distinct training contexts: all one answer in SAME8 (8:0), or four per answer in CROSS8 (4:4). Recall queries the recorded symbol for all 384 names. The fully supplied (EXPLICIT) and omitted-input (DIRECT) routes use the same 768 test cases: 96 trained names × eight held-out contexts. They are different questions, not independent replications.

| | Record recall | | All six inputs supplied | | Recorded input omitted | |
|---|---|---|---|---|---|---|
| Block | SAME8 | CROSS8 | SAME8 | CROSS8 | SAME8 | CROSS8 |
| Qwen W3 | 100.00 | 100.00 | 97.79 | 97.66 | 47.01 | 88.54 |
| Qwen W1 | 100.00 | 98.70 | 97.01 | 96.48 | 56.25 | 87.37 |
| Qwen W4 | 100.00 | 100.00 | 97.01 | 97.01 | 47.92 | 85.81 |
| Llama W3 | 100.00 | 100.00 | 96.48 | 95.70 | 49.87 | 88.28 |
| Phi R1 | 100.00 | 100.00 | 97.14 | 95.18 | 41.67 | 80.21 |
| Phi R2 | 100.00 | 100.00 | 94.66 | 94.14 | 41.15 | 82.29 |

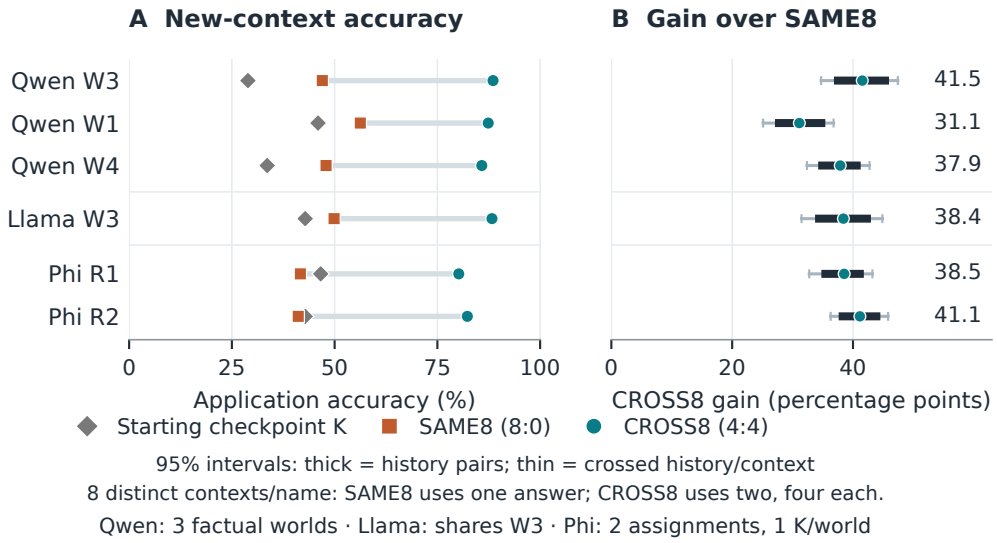



Figure 2: Omitted-input application on 768 test cases per comparison (96 trained names, eight contexts each), held out from application training. K is the starting checkpoint; SAME8/CROSS8 denote 8:0/4:4 allocations. Left: absolute accuracies. Right: paired percentage-point gains. Thick intervals resample histories; thin intervals also reweight contexts. Phi assignments share K and W3; Llama also uses W3. Acquisition differs across families, precluding an architecture ranking.

**A large benefit appears before answers are balanced.** Figure 3 shows the same trained-name test under intermediate assignments. Moving from 8:0 to 7:1 raises W3 accuracy from 47.01% to 82.68%. W1 rises from 56.25% to 85.42%, and W4 from 47.92% to 83.33%. All three gains remain positive under the nine-comparison adjustment, both uncertainty schemes and batch-size-one scoring. Each also improves above K. The 6:2 changes are smaller and heterogeneous. Exact balance is thus unnecessary for a large benefit here, but neither an optimal allocation nor general equivalence to 4:4 is established. One minority context still means six presentations, not one exposure.

**Equivalent records yield more consistently correct answers.** With the same record and context, changing the name should not change the correct category. Both names are answered correctly more often under CROSS8: Phi's first assignment rises from 1.56% to 69.27%. Distributional differences also decrease, so the change is not merely more consistent wrong answers. Full both-correct and total-variation results are in Table 5. Phi's CROSS8 distributions nevertheless differ more across names than K's, another reason to keep the starting-model comparison.

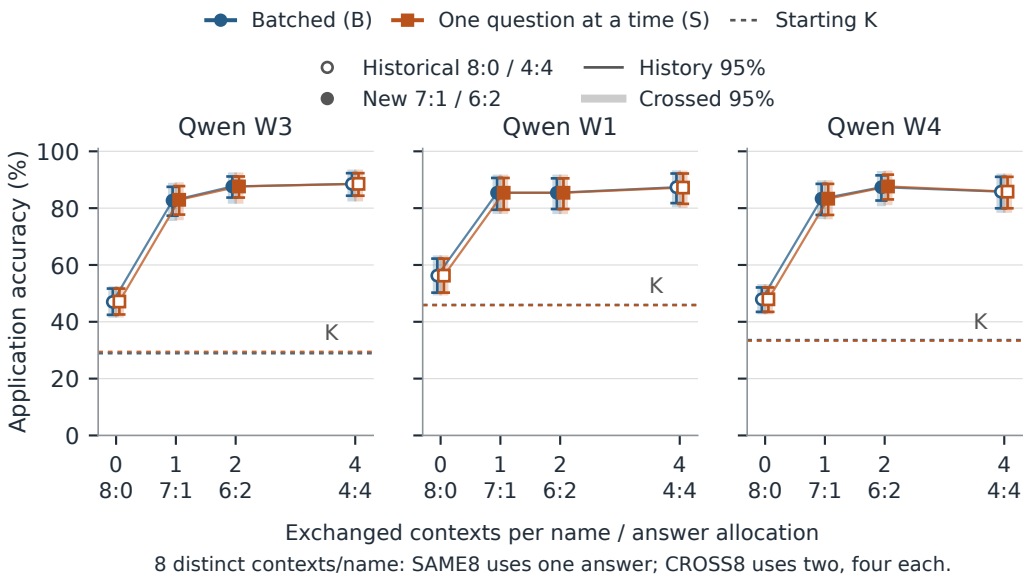



Figure 3: Graded allocations on the same 768 trained-name test cases in each Qwen world. Batched (B) and one-question-at-a-time (S) measurements use identical weights. Historical 8:0/4:4 and new 7:1/6:2 models are distinguished; starting checkpoint K is separate. Intervals are the extension's conditional history and crossed 95%. Exchange positions are 0, 1, 2 and 4; connecting segments are not a fitted law. The 7:1 minority context is presented six times.

**Could application labels themselves reveal the record?** Indirect learning is possible (Treutlein et al., 2024). We calculate which candidate symbols remain consistent with the eight labeled contexts for each name. This tests what the training examples determine, not what the model inferred. CROSS8 uniquely identifies the symbol for every trained name; SAME8 does so for 73–85 of 96 names. Restricting to names identifiable under both conditions retains positive CROSS8 effects in every original comparison, from 32.16 to 47.92 percentage points under both interval schemes. Complete original pairs give the same qualitative result (Tables 8–10).

The graded 7:1 advantage also survives a separate restriction requiring identification under all four application allocations, and its complete-pair sensitivity (Appendix H). K is not an application allocation. These post hoc restrictions narrow ambiguity as a sole explanation, but do not equalize learning difficulty or rule out reinforcement or relearning of the record.

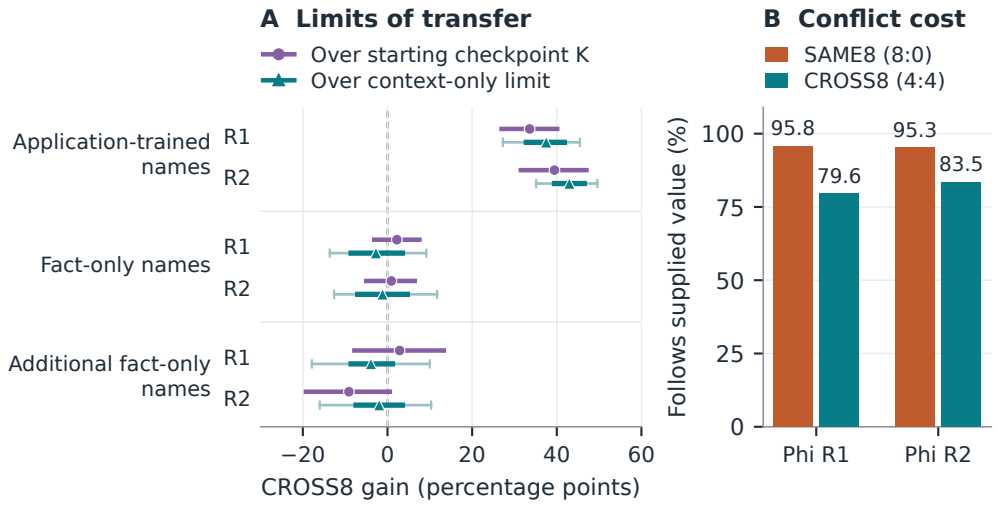
**The two transfer tests have different limits.** First, on fact-only test names, Qwen and Llama improve over both SAME8 and K. Phi improves over SAME8, but its gain over K is small and unresolved (Table 2). A sibling advantage can therefore reflect less deterioration under continued training. The additional fact-only population is separate: Phi's second assignment has a negative parent-relative point and an unresolved crossed sibling comparison. No Phi assignment clears the context-only bound on fact-only names; that bound is exceeded more consistently on application-trained names (Figure 4).

The 7:1 Qwen models also improve the original fact-only test over both 8:0 and K in all three worlds, reaching 48.31%, 57.36% and 73.89%, respectively. These are secondary unadjusted comparisons, positive under both schemes and scoring paths. In the additional fact-only population, W3's improvement over K remains unresolved.

Second, the **answer-unseen** test concerns names that *did* receive application examples. It retains targets absent from the union of that name's original 8:0 and 4:4 training labels. For a name taught only $A$ under 8:0 and $A/B$ under 4:4, a test requiring $C$ uses a previously unpaired name–answer combination. Labels remain familiar globally. This post hoc subset contains only recorded symbols $P$ and $S$, with 106–156 cases and 17–26 histories per comparison.

Table 2: Application with the recorded input omitted on fact-only TEST names: 192 names and eight contexts, giving 1,536 cases per comparison. K is the starting checkpoint; S is SAME8 (8:0), C is CROSS8 (4:4). Accuracies are percentages and differences are percentage points. Brackets are history-pair 95% intervals. Additional fact-only names and crossed sensitivities remain separate in Appendix D.

| Block | K | S | C | $C - S$ [95%] | $C - K$ |
|---|---|---|---|---|---|
| Qwen W3 | 29.88 | 41.41 | 54.30 | 12.89 [7.88,17.97] | 24.41 |
| Qwen W1 | 45.44 | 46.61 | 64.19 | 17.58 [12.89,22.59] | 18.75 |
| Qwen W4 | 37.63 | 55.53 | 69.60 | 14.06 [9.96,17.77] | 31.97 |
| Llama W3 | 34.96 | 53.84 | 70.57 | 16.73 [12.17,21.55] | 35.61 |
| Phi R1 | 45.96 | 38.80 | 48.18 | 9.38 [4.88,14.06] | 2.21 |
| Phi R2 | 43.82 | 29.95 | 44.73 | 14.78 [9.24,20.90] | 0.91 |




A: 95% intervals on two distinct evaluation sets; see caption.
B: descriptive rates on 768 supplied-conflict cases per checkpoint.
8 distinct contexts/name: SAME8 = one answer; CROSS8 = two answers, four each.

Figure 4: Phi's transfer and conflicting-input limits. Purple: ordinary accuracy change over starting checkpoint K (history 95%). Teal: margin above the context-only bound on a separate balanced grid (context-primary thick, crossed thin). These use different supports and estimators. Additional fact-only names retain the legacy identifier "unexposed". Conflict bars are descriptive supplied-target accuracy on 768 trained-name cases per checkpoint. SAME8/CROSS8 are the 8:0/4:4 allocations.

CROSS8 exceeds SAME8 under both schemes in five of six answer-unseen comparisons, but none establishes a gain over K. Both Phi assignments are below K. Thus the resolved sibling advantage extends beyond selecting a name's familiar application labels without establishing parent-relative improvement on this restriction (Tables 32 and 33).

**Better application can coexist with worse use of supplied inputs.** In the conflict test, the supplied symbol is the value the model is instructed to use. Phi's first assignment falls from 95.83% accuracy under SAME8 to 79.56% under CROSS8; the second falls from 95.31% to 83.46%. Answers consistent with the original record become more frequent. These descriptive fixed-panel rates show a cost, not a necessary trade-off or an identified memory-selection process.

The graded study finds a Qwen W4 cost too: 95.31% under 8:0 falls to 80.08% under 7:1, a $-15.23$-point difference with crossed 95% interval $[-23.01, -8.43]$. Scoring one question at a time retains the loss; W3 and W1 do not show the same resolved effect. All six new models have perfect

measured primary recall, but the full $\pm 3$-point component-compatibility requirement is not established. It requires the paired intervals to lie within the margin in every governing cell (Appendix H).

## 5 INTERNAL INTERVENTIONS

Can selected learned computations transfer part of the original SAME8/CROSS8 difference? We replace a feed-forward component from SAME8 with its CROSS8 counterpart. Such a component, an MLP, transforms each token's internal representation. The complete imported function acts on the recipient's current input to that layer; external normalization remains recipient-native. No donor-question activation or missing fact is inserted. The temporary combined model is the **hybrid**, distinct from cached-activation replacement (Biran et al., 2024; Dai et al., 2026).

Qwen layers 12–17 were selected on W3 development examples and fixed for W1/W4; Llama uses 9–13. Separate evaluations contain 384 and 768 cases with different numerical contracts; they are not pooled. Effects are partial, with Llama's 768-case effect unresolved. Starting-checkpoint donors, alternative locations and reciprocal replacements show that CROSS8's functions and the selected location are not uniformly superior (Tables 11–12).

The **prediction position** is the final non-padding input token from which the next answer is scored. **Prefix** replacement acts at the other real input positions. W3 gains 69 correct cases from prefix replacement, compared with 72 at all positions and two at the prediction position alone. W1 and W4 show the same pattern; closeness is not equivalence or evidence of retrieval. Separately, fitted constant output-score shifts fail the declared fidelity criteria. Summing input-dependent changes from individual replacements meets the additive criteria in W4 with the CROSS8 donor, but not W3. These tests constrain those particular explanations (Figure 5, Appendix G.1).

Finally, we restore a later component's output to its saved value in the unmodified SAME8 model. We measure replacement with and without restoration, and apply the same restoration to the unmodified model as a control. This four-condition comparison tests dependence within the intervention. Selected Qwen components reduce the hybrid benefit; other effects are heterogeneous, probability-only or sparse (Table 7). It does not establish natural necessity or a complete causal abstraction (Geiger et al., 2025; Makelov et al., 2024). These previously inspected interventions concern the original 8:0/4:4 models, not the mechanism of the new 7:1 result.

## 6 RELATED WORK

**Recall versus application.** Correct factual answers need not support successful composition. Press et al. (2023) compare component questions with multi-hop questions, which require multiple linked facts. Allen-Zhu & Li (2025) study retrieval and manipulation of synthetic biographies, including individuals represented in factual training but not in manipulation training. Dai et al. (2026) fine-tune on single facts and test chaining and intersection questions. Their self-patching replaces an internal activation during a forward pass with a cached value, including from the same prompt, to investigate knowledge–circuit misalignment. Our experiment instead changes how application examples are assigned to names while matching factual supervision and overall task material. It does not directly test that misalignment account.

**Training organization and supervision.** Lampinen et al. (2025) compare in-context learning and fine-tuning on controlled holdouts. They examine inference-based augmentation and sentence splitting that changes training correlations, accounting for dataset size and gradient steps. Splitting has condition-dependent effects. Akyürek et al. (2024) generate implications and contradictions, then train on consistency-selected statements. Chaudhry et al. (2026) combine reasoning-trace supervision with correctness-feedback reinforcement learning and test knowledge excluded from that reasoning training. We reassign an existing application pool between names with the same recorded symbol. This preserves overall context–recorded-symbol–answer combinations, eight contexts per name, auxiliary examples and updates. The experimental factor differs from sentence grouping, inference enrichment and reasoning training; performance is not compared across these different tasks.

**Compositional generalization and shortcuts.** Systematic generalization concerns recombining familiar components (Lake & Baroni, 2018). Natural-language inference judges relations such as entailment between sentences. Controlled examples expose syntactic shortcuts in this task (McCoy et al., 2019). Human-authored counterfactual revisions, including equal-size comparisons, test reliance on spurious training features (Kaushik et al., 2020). Yang et al. (2025b) constrain shortcut opportunities in latent multi-hop evaluation. Our 8:0 assignment permits a constant answer associated with each name, a shortcut opportunity in the sense of Geirhos et al. (2020). The 4:4 assignment changes that dependency while retaining correct labels and matched material. Several joint dependencies change, so answer entropy is not isolated as the cause.

**Indirect learning and conflicting inputs.** Treutlein et al. (2024) show that models can infer latent information from indirect training examples. Application labels could similarly reveal our omitted recorded symbol. Our identifiability audit measures what those labels determine, not what the model learned. Knowledge-conflict studies examine competition between stored information and supplied context (Longpre et al., 2021), and prompting for context faithfulness (Zhou et al., 2023). Our conflicting-input test asks whether improved application also supports following a conflicting supplied value.

**Internal interventions.** Yang et al. (2024) distinguish intermediate recall from its use, while Biran et al. (2024) examine delayed intermediate resolution through activation interventions. Model stitching connects frozen network portions with an adapter, a learned connecting transformation (Bansal et al., 2021). Our replacement evaluates a donor's learned function on the recipient's current internal input, without fitting a connector or copying a donor-question activation. Patching conclusions depend on measurement choices (Zhang & Nanda, 2024), and localization need not predict edit efficacy (Hase et al., 2023). Donor, location and reciprocal controls therefore qualify our functional effects without establishing a unique natural computation.

## 7 Discussion and limitations

Assignment changes which answers can be predicted from a name alone. A large benefit appears before balance in this task, without establishing a universal exception threshold or an optimal allocation.

Application labels may reinforce a record or teach a name-specific computation. Identifiability does not establish equal learning difficulty or unchanged representations. New contexts, unpaired answers, fact-only names and conflicting supplied inputs remain distinct tests; internal interventions do not identify a unique algorithm.

Scope is limited to one finite task, three Qwen worlds, one Llama comparison and two Phi assignments sharing a world and K. Historical exposure, prior inspection, differing acquisition and post-failure numerical amendments remain limitations. Neither competing dated updates nor general natural-language transfer was tested.

## 8 Conclusion

The tested models receive matched factual supervision, overall application material, context counts and updates, yet their ability to apply records depends strongly on which contexts accompany each name. Requiring different answers improves performance on new contexts. One distinct minority-answer context, presented six times, already yields a large benefit in three Qwen worlds; exact balance is unnecessary for that benefit.

The improvement is not universal. Beating a weaker continuation need not mean improving above the starting model, and ordinary application can improve while compliance with supplied inputs worsens. Internal replacements partly change the original behavior without identifying a unique computation. The experiment therefore distinguishes receiving factual information from the uses of that information supported by its training examples.

### AI USE STATEMENT

Generative-AI tools assisted with implementation, debugging, execution orchestration, provenance audits, result interpretation, literature discovery, figure preparation and manuscript drafting. Synthetic examples and labels were constructed by explicit deterministic programs and the HMM oracle. Reported observations come from executed model evaluations. Source identities, per-item outputs, fixed comparison definitions and numerical checks provide an audit trail for AI-assisted work. The authors take responsibility for the final manuscript, scientific claims, and accompanying artifacts, including all AI-assisted contributions.

### ETHICS STATEMENT

Entities and observations are fictional; the task does not involve human participants or real personal records. It concerns deliberately introduced post-training information, not control over unknown pretraining provenance or reliable behavior in consequential applications. Better ordinary factual use can coexist with worse compliance with conflicting inputs, so a primary-benchmark improvement should not be read as uniformly safer behavior.

### REPRODUCIBILITY STATEMENT

The appendices describe the task, training, populations and generation-specific estimators. The scientific supplement preserves the historical release and adds the graded-extension module with full-precision results, assignment and membership records, checkpoint identities and recorded observations. Its README distinguishes display regeneration, the historical identifiability analysis and both complete graded-extension statistical analyses from supplied predictions. Clean-copy commands and their scoped verification receipts are provided. These checks do not constitute whole-study neural reproduction; model execution requires the external weights and numerical environment listed in the omission records. Appendix H.2 preserves the extension's failed original numerical gates and amended measurement policy.

## REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025. doi: 10.48550/arXiv.2503.01743. URL https://arxiv.org/abs/2503.01743.

Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Wijaya, and Jacob Andreas. Deductive closure training of language models for coherence, accuracy, and updatability. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9802–9818. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.584. URL https://aclanthology.org/2024.findings-acl.584/.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d5494c8747276d3cdb2598e5617de89d-Abstract-Conference.html.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14113–14130. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.781. URL https://aclanthology.org/2024.emnlp-main.781/.

Arslan Chaudhry, Sridhar Thiagarajan, and Andrew Lampinen. Improving latent generalization using test-time compute. arXiv preprint arXiv:2604.01430, 2026. URL https://arxiv.org/abs/2604.01430v1. Version 1, April 1, 2026.

Lu Dai, Ziyang Rao, Yili Wang, Hanqing Wang, Hao Liu, and Hui Xiong. Towards mechanistically understanding why memorized knowledge fails to generalize in large language model finetuning. arXiv preprint arXiv:2607.08393, 2026. URL https://arxiv.org/abs/2607.08393v1. Version 1, July 9, 2026.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://www.jmlr.org/papers/v26/23-0058.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0774. URL https://papers.nips.cc/paper_files/paper/2023/hash/3927bbdcf0e8d1fa8aa23c26f358a281-Abstract-Conference.html.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882. PMLR, 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Andrew K. Lampinen, Arslan Chaudhry, Stephanie C. Y. Chan, Cody Wild, Diane Wan, Alex Ku, Jörg Bornschein, Razvan Pascanu, Murray Shanahan, and James L. McClelland. On the generalization of language models from in-context learning and finetuning: A controlled study. arXiv preprint arXiv:2505.00661, 2025. URL https://arxiv.org/abs/2505.00661v3. FoRLM Workshop at NeurIPS 2025; arXiv version 3, November 10, 2025.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/70b8505ac79e3e131756f793cd80eb8d-Abstract-Conference.html.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Meta. Llama 3.2 model card. Official model documentation, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct. Released September 25, 2024; accessed September 24, 2026.

Art B. Owen and Dean Eckles. Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics*, 6(3):895–927, 2012. doi: 10.1214/12-AOAS547. URL https://doi.org/10.1214/12-AOAS547.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Johannes Treutlein, Dami Choi, Jan Betley, Sam Marks, Cem Anil, Roger Grosse, and Owain Evans. Connecting the dots: LLMs can infer and verbalize latent structure from disparate training data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4465. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fe489a28a54583ee802b8e2955c024c2-Abstract-Conference.html.

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

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10210–10229. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.550. URL https://aclanthology.org/2024.acl-long.550/.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3971–3992. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.findings-acl.205. URL https://aclanthology.org/2025.findings-acl.205/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.968. URL https://aclanthology.org/2023.findings-emnlp.968/.

# A TASK, POPULATIONS, AND TRAINING CONSTRUCTION

**Exact oracle.** Let $o_{1:6} = (c_1, \ldots, c_5, v_e)$, where $v_e \in \{P, Q, R, S\}$ is the recorded endpoint for entity $e$. The two latent states have initial probability $\pi = (1/2, 1/2)$, transition matrix $T$, and emission matrix $E$ (columns ordered $P, Q, R, S$):

$$T = \begin{pmatrix} .85 & .15 \\ .15 & .85 \end{pmatrix}, \qquad E = \begin{pmatrix} .15 & .45 & .10 & .30 \\ .30 & .10 & .45 & .15 \end{pmatrix}. \tag{2}$$

The reference uses float64 filtering. Initialize $\widetilde{\alpha}_1(j) = \pi_j E_{j,o_1}$ and normalize it as below. For $t > 1$, update and normalize:

$$\widetilde{\alpha}_t(j) = E_{j,o_t} \sum_i \alpha_{t-1}(i) T_{ij}, \qquad \alpha_t(j) = \frac{\widetilde{\alpha}_t(j)}{\sum_k \widetilde{\alpha}_t(k)}. \tag{3}$$

The target is $F(c, v_e) = \mathrm{bin}(\alpha_6(1))$, with bins A, B, C, D corresponding to $[0, .25)$, $[.25, .5)$, $[.5, .75)$, and $[.75, 1]$. Each target is computed by this oracle; the model is not asked to produce a numerical posterior or a chain of thought. Reachable bins depend on the endpoint: Q and R cannot reach all four categories. The construction chooses two bins with sufficient eligible support rather than imposing impossible label balance.

**The date denotes the lookup key, not a temporal-update treatment.** The five supplied observations are hypothetical inputs; the sixth comes from a fixed dated record. The archived Qwen/Llama and Phi primary application prompts explicitly direct the model to use the supplied five observations instead of the original prefix. A concrete archived request is illustrated below with line breaks compacted for display. The renderer passes a user message, without a task-specific system definition; native chat delimiters and any generic template header are retained in the machine-readable witnesses. No endpoint, target, treatment name or anchor label is supplied in this DIRECT request.

> For this hypothetical calculation, use the five supplied signals rather than the entity's recorded earlier signals. Only the missing 2020 endpoint comes from the archive.
>
> `METHOD: RISK-v1; TARGET_YEAR: 2020`
> `EVIDENCE_AS_OF: 2020-12-31; END_BOUNDARY`
> `ENTITY: Analy Lazy`
> `OBSERVATIONS:`
> `2015: S; 2016: Q; 2017: Q; 2018: P; 2019: Q`
> `2020: ARCHIVE(Analy Lazy, 2020)`
>
> Resolve the archive entry for the requested year. Apply RISK-v1 to the observations through that year. Return one risk-bin label: A, B, C, or D.
> `ANSWER:`

For this entity, the archived factual query is `ENTITY: Analy Lazy`, `RECORD_YEAR: 2020`, followed by "Return the signal recorded for this entity in that year: P, Q, R, or S." and `ANSWER:`; its target is $Q$. The explicit application version replaces the archive reference by `2020: Q`. Independent filtering of $SQQPQQ$ gives $q = 0.06335134344541139$, so its application target is $A$. These oracle answers do not assert how a particular checkpoint responded. The example is selected by the smallest eligible archived case identifier, independently of its scored outcome. Eighteen original-comparison route witnesses (three routes in each of six blocks) reproduce the recorded prompt-token hashes from local tokenizers. This experiment does not add newer conflicting records and does not test selective control over competing vintages of knowledge.

**Entity partitions and aliases.** The 384 entities form 192 original pairs sharing the relevant endpoint and original history prefix. Application training uses 96 entities/48 pairs; factual-only validation uses 48/24; factual-only test uses 192/96; a separately retained factual-only population uses 48/24. The latter's legacy name, `FACT_ONLY_UNEXPOSED`, refers to an earlier future-update allocation, not absence of historical fact learning or unfamiliar wording. We keep it separate. The primary ordinary application tests use eight contexts per entity: 768, 1,536, and 384 cases for trained, factual-only test, and the additional factual-only populations, respectively. P1, P3, and P3U name these three contrasts; P2 is same-fact alias total variation on a common context grid.

Table 3: Context-pool sizes after original-history exclusion and measured exposure of the historical K parent. Llama shares the Qwen W3 context set. Phi acquires a new parent and does not inherit these exposure counts.

| World | Train | Validation | Test | Test seen by old K |
|---|---|---|---|---|
| Qwen W3 | 512 | 173 | 161 | 105/161 (65.22%) |
| Qwen W1 | 495 | 176 | 174 | 109/174 (62.64%) |
| Qwen W4 | 515 | 172 | 162 | 98/162 (60.49%) |

**Matched reassignment.** For each application-training pair $(e_0, e_1)$ with endpoint $v$, choose eight distinct contexts $C_a$ yielding bin $a$ and eight $C_b$ yielding bin $b \neq a$. SAME8 assigns all of $C_a$ to $e_0$ and $C_b$ to $e_1$. CROSS8 retains those assignments for indices 0–3 and exchanges names for indices 4–7. Every name therefore has eight distinct contexts, with one constant target or two targets presented four times each. The two paired composite examples enter the same effective optimizer batch, along with four shared generic explicit examples and two shared factual examples. Name, context, target, and auxiliary-example marginals match at every effective step; joint name–context assignments intentionally differ. Three training renderings and two macrocycles give 48 composite presentations per trained entity.

**Training-label entropy and naming.** The empirical label entropies are $H_{\mathrm{SAME8}}(Y \mid e) = 0$ and $H_{\mathrm{CROSS8}}(Y \mid e) = 1$ bit. These are properties of constructed training labels, not model uncertainty. The reassignment also changes name–context associations and which recorded symbols those labels identify. The ratio names 8:0/7:1/6:2/4:4 correspond to raw condition IDs M0/M1/M2/M4. In older tables, T/F/U denote trained, fact-only and additional fact-only populations; the graded-extension view `U` instead denotes answer-unseen cases. The raw population key `FACT_ONLY_UNEXPOSED` never means an absence of factual training.

**Context exposure.** A fixed permutation of the $4^5 = 1,024$ possible contexts assigns 614 to training, 205 to validation, and 205 to test. Removing each world's original recorded prefixes gives Table 3. Child composite and generic training use only the new training pool. However, inherited component parents had already seen many of these test strings as fully explicit problems. Thus "held out" means held out from this child curriculum; it does not mean a globally novel string or an untouched research programme. Llama and Phi use world-3 semantics; Phi's second assignment changes the allocation and optimizer seed, not the factual world.

## B PARENT ACQUISITION, OPTIMIZATION, AND NUMERICAL EXECUTION

**Historical parents.** Qwen3-4B and Llama-3.2-3B-Instruct use previously acquired factual H checkpoints, followed by shared component-only K preparation. Existing auxiliary facts in H are retained; we do not claim that their entire acquisition history comprises only the 384 endpoint records. The older supplementary entity-composite curriculum is omitted from the new controlled treatment. The audited K streams contain 75% generic explicit examples and 25% factual replay, with zero composite examples: 10,552 rows for Qwen W3/W4, 10,560 for W1, and 10,248 for Llama. These are separate inherited parents, not identically acquired roots across families. Each paired treatment starts from the same within-block K weights; historical and repeated assignments within a world are not independent factual-world replications.

**Prospective Phi acquisition.** The root is `microsoft/Phi-4-mini-instruct`, revision `cfbefacb99257ffa30c83adab238a50856ac3083`; all 3,836,021,760 unique language-model parameters are trainable. Its native implementation uses the Phi3 model class, which does not indicate substitution of a Phi3 checkpoint. New H training uses 384 names $\times$ three factual renderings per epoch. Checks at epochs 8/16/32/64 select the earliest passing checkpoint: overall recall at least 98%, each original population 95%, worst endpoint 90%, and validity 98%. Epoch 8 passes with 9,216 examples and 1,152 updates. Fixed paraphrases are descriptive, not additional selection gates.

K receives six generic explicit and two factual examples per effective batch. Checks at 512/1,024/2,048 updates require factual recall 95%, worst-endpoint recall 80%, generic accuracy 85%, and validity 98%. K512 fails the generic threshold (82.23%); K1024 passes (92.34%, with 100% measured recall). It processes 8,192 examples and 982,550 non-padding tokens. No DIRECT performance criterion selects H or K. Both prescribed Phi assignments then branch from this same K, irrespective of the first treatment result. They are two assignment/optimization repetitions conditional on one parent and one world.

**Fixed child recipe.** The twelve original SAME8/CROSS8 children (six sibling pairs) use full-parameter BF16 training with AdamW8bit: learning rate $2 \times 10^{-5}$, betas $(.9, .999)$, epsilon $10^{-8}$, zero weight decay, gradient norm clip 1, 50-update warmup, then constant learning rate. Microbatch 4 and accumulation 2 give effective batch 8. Each original child receives 2,304 updates and 18,432 rows: 4,608 composite, 9,216 generic explicit, 4,608 factual replay. The prompt is masked; the complete answer plus inherited end marker is supervised, with per-example token mean followed by effective-batch mean. Fresh optimizer states and equal seeds within siblings separate arms. The block seeds are 61003/61001/61004 for Qwen W3/W1/W4, 61103 for Llama, and 61431/61432 for Phi. Execution order is counterbalanced.

Snapshots at 0.5 and 1 macrocycle provide validation trajectories; the fixed 2-cycle checkpoint is always primary. There is no selection of the most favorable validation child. Actual receipts confirm the declared row/update count for all twelve original children. Non-padding input tokens per original child are 2,781,696 for Qwen, 2,864,664 for Llama, and 2,244,156 for Phi; every original child has 36,864 supervised answer/end-marker tokens. Within pairs, these totals and per-step target-suffix multisets match. Phi padding totals differ slightly between arms. Thus the matched budget concerns semantic presentations, supervised content, and optimizer updates, not identical cross-family tokens or FLOPs. The six intermediate-allocation Qwen children and their execution records are described in Appendix H.

**Scoring and repairs.** The primary evaluation uses unrestricted full-vocabulary top-1 at the answer boundary. Outputs outside the four atomic valid answers are wrong. Distributional diagnostics retain full-vocabulary label probabilities plus aggregate outside mass; primary results do not renormalize only the valid labels. Answer-token identities and native chat templates are checked for each model. By contrast with the supplementary route below, the reported DIRECT panels retain every fixed case.

**Supplementary recall-then-classify route (SELF).** SELF first queries the recorded symbol and uses that model's corresponding cached explicit response with its returned symbol supplied. A valid recall is an allowed $P, Q, R, S$ output, not necessarily a correct one. The paired primary includes cases with valid recall in *both* compared models. Even after a wrong valid recall, classification is judged against the original target, not a target redefined by that recalled symbol. Invalid retrieval has no fabricated second-stage answer. Because the paired-valid selection is post-treatment, an all-case sensitivity counts invalid recalls as wrong. This route is a composition of recorded recall and explicit responses, not evidence of the model's natural internal reasoning sequence.

Phi uniformly uses BF16 singleton scoring after a batched-versus-singleton native-generation check disagreed on one acquisition item. The same singleton convention applies to all retained checkpoints and routes; this is a numerical repair, not outcome-specific row substitution. The earlier failed batch scoring and fixed thresholds are preserved in provenance. Historical Qwen/Llama use their qualified native scoring paths; function-transfer evaluations use BF16 and explanatory interventions use separately qualified FP32. Cross-generation panel and numerical differences preclude treating their effects as a precision-only replication.

## C ESTIMANDS AND UNCERTAINTY

For an ordinary DIRECT panel, define $d_i = \mathbf{1}[\text{CROSS8}(x_i) = y_i] - \mathbf{1}[\text{SAME8}(x_i) = y_i]$. We first average within the original history-pair cluster, then average pair means. The primary 95% interval resamples these pair means 2,000 times with replacement (declared seed 61991). The crossed sensitivity independently draws Poisson(1) weights for history pairs and visible contexts and multiplies their weights for each case. Percentile intervals describe uncertainty conditional on the fitted mod-

els and fixed panel construction; cases are not additional training replications. Tables retain both conventions even when their conclusions differ.

For background on product reweighting over crossed factors, see Owen & Eckles (2012). Their analysis includes asymptotic extensions under additional assumptions. It does not by itself establish finite-sample coverage for the nonlinear paired statistics and conditional panels here. The executed weights, normalization and percentile rules remain those stated in this appendix.

For the three Qwen worlds, the primary summary is the equal-world mean. The retained df $= 2$ t interval uses sample standard deviation across the three world effects and multiplier 4.303. Llama shares W3 semantics and is not a fourth Qwen world. Phi's two assignments have a descriptive mean and range, without a narrow pooled normal/t interval. The final mean differences are 36.8489583 pp (Qwen), 38.4114583 pp (Llama), and 39.84375 pp (Phi); one-decimal reporting gives 36.8, 38.4, and 39.8.

**Archived original headline contrasts.** The original trained-name Qwen equal-world gain is 36.8 pp with world interval $[23.72, 49.98]$; its three effects are 41.54, 31.12 and 37.89 pp. Llama gains 38.41 pp. Phi gains 38.54 and 41.15 pp, with history intervals $[34.77, 41.80]$ and $[37.63, 44.53]$ and positive crossed sensitivities. These quantities concern the original SAME8/CROSS8 comparisons, not means over intermediate allocation levels.

**Alias and factual-dependence diagnostics.** Alias TV is $\frac{1}{2}\sum_{z\in\{A,B,C,D,\text{outside}\}} |p(z \mid e_0, c) - p(z \mid e_1, c)|$, with SAME8$-$CROSS8 orientation (opposite the accuracy contrast). Both-aliases-correct and valid mass rule out interpreting agreement alone as success. On the endpoint-balanced lattice, the context-only bound is

$$B_0 = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \max_y \frac{1}{4} \sum_{v\in\{P,Q,R,S\}} \mathbf{1}[F(c,v) = y]. \tag{4}$$

Each arm's lattice accuracy is compared with this bound on the same balanced support, with contexts as the outer resampling unit and a crossed sensitivity. The ordinary panel has different weighting: its accuracy must not be subtracted from this $B_0$. Clearing $B_0$ excludes a strictly context-only predictor on that panel; it does not establish use of an unchanged pre-existing fact representation. Composite supervision itself may teach or reveal the endpoint of trained entities. Fact-only transfer is a stronger, separately reported constraint.

**Component compatibility and mechanistic intervals.** The original component criterion is two-sided compatibility within 3 pp on every applicable validation/test partition. A nonsignificant difference is not equivalence; ceiling recall cannot establish equality of internal knowledge. The function-transfer study's primary function/reader intervals use crossed history/context Exp(1) multiplier weights (2,000 draws), once per unique cluster per row. The explanatory-intervention study's primary interval uses history multipliers and reports crossed history/context sensitivity. Declared near-zero sign checks use more draws where required. These are conditional intervention-panel intervals, not confidence intervals over architectures or circuit discovery. The Qwen world-mean function-transfer effect is 5.46875 pp, with df 2 interval $[-6.6511, 17.5886]$ pp; Q3 was the selection world. Related mechanistic panels are not independent observations of a common training effect.

## D SUPPLEMENTARY RESULTS AND QUALIFICATIONS

Table 4 separates parent K from both children. Every trained-entity contrast includes a substantial improvement of CROSS8 over K, not merely deterioration of SAME8. This does not hold uniformly for factual-only transfer. In Phi, CROSS8$-$K on F is+2.21 pp (R1) and+0.91 (R2); on U it is+2.86 and$-9.11$. All four intervals include zero. The relative sibling advantage therefore coexists with unresolved gains above the parent and larger SAME8 losses. Likewise, Phi CROSS8 clears the balanced context-only bound on trained entities, but its F/U margins remain unresolved. Table 5 reports agreement jointly with successful answering.

Table 4: Complete ordinary behavioral levels and contrasts. K/S/C are accuracies in percent; differences and intervals are percentage points. T: application-trained; F: factual-only test; U: separate legacy factual-only population. Paired and crossed intervals are retained separately.

| Block | Pop. | K | S | C | C−S | Pair 95% | Crossed 95% |
|---|---|---|---|---|---|---|---|
| Qwen W3 | T | 28.91 | 47.01 | 88.54 | 41.54 | [36.85, 45.96] | [34.72, 47.42] |
| Qwen W3 | F | 29.88 | 41.41 | 54.30 | 12.89 | [7.88, 17.97] | [6.12, 19.73] |
| Qwen W3 | U | 30.21 | 35.16 | 48.70 | 13.54 | [2.86, 24.74] | [0.26, 27.87] |
| Qwen W1 | T | 45.96 | 56.25 | 87.37 | 31.12 | [27.08, 35.42] | [25.12, 36.78] |
| Qwen W1 | F | 45.44 | 46.61 | 64.19 | 17.58 | [12.89, 22.59] | [10.92, 24.93] |
| Qwen W1 | U | 38.02 | 48.96 | 67.97 | 19.01 | [8.59, 30.47] | [5.88, 33.15] |
| Qwen W4 | T | 33.59 | 47.92 | 85.81 | 37.89 | [34.24, 41.28] | [32.39, 42.74] |
| Qwen W4 | F | 37.63 | 55.53 | 69.60 | 14.06 | [9.96, 17.77] | [8.15, 20.08] |
| Qwen W4 | U | 40.62 | 57.55 | 66.67 | 9.11 | [-0.52, 20.31] | [-4.65, 21.95] |
| Llama W3 | T | 42.84 | 49.87 | 88.28 | 38.41 | [33.72, 42.97] | [31.49, 44.86] |
| Llama W3 | F | 34.96 | 53.84 | 70.57 | 16.73 | [12.17, 21.55] | [10.36, 23.77] |
| Llama W3 | U | 34.38 | 49.22 | 71.61 | 22.40 | [13.54, 31.26] | [8.43, 36.90] |
| Phi R1 | T | 46.61 | 41.67 | 80.21 | 38.54 | [34.77, 41.80] | [32.76, 43.22] |
| Phi R1 | F | 45.96 | 38.80 | 48.18 | 9.38 | [4.88, 14.06] | [3.07, 15.32] |
| Phi R1 | U | 51.30 | 36.20 | 54.17 | 17.97 | [7.81, 28.39] | [4.52, 32.08] |
| Phi R2 | T | 42.84 | 41.15 | 82.29 | 41.15 | [37.63, 44.53] | [36.31, 45.82] |
| Phi R2 | F | 43.82 | 29.95 | 44.73 | 14.78 | [9.24, 20.90] | [7.21, 22.88] |
| Phi R2 | U | 48.96 | 28.12 | 39.84 | 11.72 | [2.86, 19.53] | [-0.99, 24.04] |

Table 5: Same-fact alias diagnostics on the trained-entity test common grid. TV is on its native 0–1 scale, not accuracy points. Both-correct rates are percentages.

| | Alias TV | | Both aliases correct (%) | |
|---|---|---|---|---|
| Block | SAME8 | CROSS8 | SAME8 | CROSS8 |
| Qwen W3 | 0.889 | 0.122 | 8.33 | 78.26 |
| Qwen W1 | 0.716 | 0.067 | 25.39 | 82.29 |
| Qwen W4 | 0.967 | 0.103 | 1.95 | 78.26 |
| Llama W3 | 0.935 | 0.115 | 6.38 | 71.35 |
| Phi R1 | 0.971 | 0.098 | 1.56 | 69.27 |
| Phi R2 | 0.975 | 0.107 | 2.08 | 77.60 |

**Alias and balanced-grid summaries.** In the original comparisons, the Qwen mean reduction in same-fact-name total variation is 0.7598. Both-correct rates and per-world distributions remain in Table 5. CROSS8 exceeds the trained-name context-only bound under both uncertainty schemes in all six comparisons. On fact-only names, Qwen W4 and Llama clear both; W3 is unresolved and W1's primary positive margin is unresolved under crossed sensitivity. Neither Phi assignment clears the fact-only bound. The bound belongs to the fixed balanced grid, not the ordinary or answer-unseen tests.

**Component limits.** Table 6 gives explicit-task accuracy on both partitions rather than selecting the better result. Llama's validation decline is 5.21 pp with a pair interval excluding $-3$ pp; test alone hides this cost. Phi R1 also fails to establish 3-point compatibility, while Phi R2's favorable pair-bootstrap compatibility is not robust to the crossed interval. Hence the original stronger compound confirmation label remains unresolved despite positive ordinary contrasts. Factual recall is 100% in both historical siblings except Qwen W1 CROSS8 (379/384 vs 384/384); all Phi checkpoints retain 384/384 on the primary factual rendering. Such ceilings should not be treated as proof of identical representations.

**Conflicting supplied values.** The explicit override route supplies a fixed alternative endpoint whose oracle answer differs from the archived answer. Successful instruction following means answering against the supplied endpoint. For Phi R1, SAME8 answers 736/768 correctly (95.83%),

Table 6: Explicit-input task competence for application-trained entities. All rows have 768 cases and 48 original history pairs. C−S and 95% pair intervals are percentage points. Component preservation requires both partitions and the dependence sensitivity, not only the displayed point differences.

| Block | Partition | SAME8(%) | CROSS8(%) | C−S | Pair 95% |
|---|---|---|---|---|---|
| Qwen W3 | val | 95.18 | 95.31 | 0.13 | [-2.09, 2.08] |
| Qwen W3 | test | 97.79 | 97.66 | -0.13 | [-1.69, 1.43] |
| Qwen W1 | val | 98.05 | 97.01 | -1.04 | [-2.73, 0.52] |
| Qwen W1 | test | 97.01 | 96.48 | -0.52 | [-1.95, 0.78] |
| Qwen W4 | val | 94.53 | 95.83 | 1.30 | [-0.91, 3.52] |
| Qwen W4 | test | 97.01 | 97.01 | 0.00 | [-1.43, 1.30] |
| Llama W3 | val | 96.48 | 91.28 | -5.21 | [-7.42, -3.12] |
| Llama W3 | test | 96.48 | 95.70 | -0.78 | [-2.21, 0.39] |
| Phi R1 | val | 95.05 | 92.32 | -2.73 | [-4.56, -0.91] |
| Phi R1 | test | 97.14 | 95.18 | -1.95 | [-4.43, 0.13] |
| Phi R2 | val | 92.58 | 92.97 | 0.39 | [-2.34, 2.99] |
| Phi R2 | test | 94.66 | 94.14 | -0.52 | [-2.73, 1.56] |

versus 611/768 (79.56%) for CROSS8. R2 gives 732/768 (95.31%) versus 641/768 (83.46%). Meanwhile, original-archive-target accuracy increases to 12.37% and 10.16% in the two CROSS8 children. These fixed-panel proportions document a cost of the intervention, not a success at selecting temporally appropriate knowledge.

**Function replacement and downstream readers.** The selected function swap replaces complete MLP gate/up/down maps and internal biases/buffers. Each donor MLP receives the recipient's live normalized input. External normalization, embeddings, attention, final norm, and output head remain recipient-native. The primary swap applies at all non-padding positions, without a donor question, oracle value, or example-specific branch. Qwen layers 12–17 inclusive were selected on Q3 development support and transferred unchanged to Q1/Q4; the executed Llama mapping is 9–13 inclusive. The function-transfer study's treated evaluation has 48 application-trained names withheld from intervention selection, not names withheld from application training.

Function-transfer forward effects are 42, 6, 15, 7 net gains out of 384 cases for Qwen W3/W1/W4 and Llama, respectively. The explanatory-intervention panel gives 72, 13, 42, 6 out of 768 on its distinct panel. The function-transfer panel uses 24 history pairs and 8 contexts; the explanatory panel uses 24 and 16. The explanatory panel's Llama interval includes zero. Selected groups are not uniquely strongest: the separate function-transfer hash-location control exceeds the selected group in Q4/Llama, and the explanatory study's immediate-neighbor comparison is another, distinct control. Neither intervention effect divided by the clean training gap is a natural mediated fraction.

Table 7 includes all 16 declared downstream readers. With clean recipient C0, function hybrid S0, clean-output clamp R0, and hybrid plus clamp SR0, the interaction is $(S0-C0)-(SR0-R0)$. Positive values indicate reduced function effect under that clamp, conditional on this intervention regime. Matched random perturbations provide three paired specificity comparisons. Llama MLP25's probability result is resolved, but its accuracy result involves only one discordant case. Native interventions are also mostly sparse: fewer than five target gains-plus-breaks occur in 1,716/1,728 Q3, 1,725/1,728 Q1, 1,728/1,728 Q4, and 1,437/1,440 Llama cells. These counts include topology-zero cells and correlated repeated sites; they are not independent evidence of a population prevalence or absence of a mechanism.

## E EVIDENCE PROVENANCE AND REPRODUCIBILITY

The manuscript evidence ledger extracts full-precision values and exact JSON pointers from completed behavioral reports, the corrected intervention evidence, and the Phi final report. Source files have SHA-256 bindings; historical block deltas are checked against the retained pooled report and parent decompositions. Training receipts establish actual example/update counts. Separately delivered packages retain raw predictions, targets, source code, checkpoint identities, audits, and

Table 7: All declared treated-panel reader interactions. Points and crossed 95% intervals are multiplied by 100 (accuracy/probability points). Specificity asks whether the probability interaction exceeds each of three matched random controls. A displayed 0.00 can represent a small nonzero bound; sparse binary effects are not reliable population lower bounds.

| Block | Reader | Accuracy [95%] | Target probability [95%] | All 3 controls |
|---|---|---|---|---|
| Qwen W3 | MLP 34 | 0.78 [0.08, 2.46] | 1.16 [0.63, 1.95] | Yes |
| Qwen W3 | MLP 35 | -0.26 [-1.48, -0.00] | 0.50 [0.05, 1.11] | Unresolved |
| Qwen W3 | Head 35:25 | 2.08 [0.53, 4.84] | 2.50 [1.24, 4.19] | Yes |
| Qwen W3 | Head 35:31 | 0.00 [0.00, 0.00] | 0.00 [-0.01, 0.02] | Unresolved |
| Qwen W1 | MLP 34 | 0.52 [0.02, 2.08] | 0.47 [0.19, 0.98] | Yes |
| Qwen W1 | MLP 35 | 0.00 [0.00, 0.00] | 0.27 [0.13, 0.52] | Yes |
| Qwen W1 | Head 35:25 | 0.00 [0.00, 0.00] | 0.25 [0.11, 0.48] | Yes |
| Qwen W1 | Head 35:31 | 0.00 [0.00, 0.00] | 0.00 [-0.01, 0.02] | Unresolved |
| Qwen W4 | MLP 34 | 0.52 [0.02, 2.01] | 0.57 [0.27, 1.01] | Yes |
| Qwen W4 | MLP 35 | 2.60 [0.73, 5.82] | 2.03 [1.05, 3.57] | Yes |
| Qwen W4 | Head 35:25 | 0.52 [0.02, 2.01] | 0.64 [0.34, 1.11] | Yes |
| Qwen W4 | Head 35:31 | -0.26 [-1.28, -0.00] | -0.01 [-0.06, 0.04] | Unresolved |
| Llama W3 | MLP 27 | 0.26 [0.00, 1.32] | 0.32 [-0.00, 0.70] | Unresolved |
| Llama W3 | MLP 25 | 0.26 [0.00, 1.32] | 0.36 [0.08, 0.75] | Yes |
| Llama W3 | Head 26:8 | 0.00 [0.00, 0.00] | 0.04 [-0.01, 0.11] | Unresolved |
| Llama W3 | Head 26:0 | 0.00 [0.00, 0.00] | 0.00 [-0.05, 0.06] | Unresolved |

reproduction records; model weights are not embedded in the manuscript ledger. The paper tables are derived from this ledger rather than rounded screenshots.

Historical revisions, failed numerical attempts, and corrected metadata are preserved. The completed function-transfer analysis was reconciled against the explanatory study's frozen group, mapping and checkpoint identities; unchanged dependencies required an interpretation update, not an additional explanatory neural evaluation. This provenance supports traceable final analyses, but it does not retroactively create an untouched holdout: historical TEST exposure, selection on Q3, shared semantic worlds, limited parent replication, and heterogeneous upstream curricula remain limitations. The six new intermediate-allocation Qwen fine-tunes and their evaluations are reported separately in Appendix H. No Gemma, temporal-update, or general natural-language-task result is claimed.

## F Post hoc endpoint identifiability and paired outcomes

We conducted a CPU-only analysis of recorded assignments and final predictions. Its reduction protocol was hashed before new stratum outcomes were examined. The original overall results and the external review's proposed count table were already known. No new model outputs were generated, and this is not prospective registration of the original experiment.

For each application-trained name and arm $a$, define

$$\mathcal{V}_e^a = \{v \in \{P, Q, R, S\} : F(c_i, v) = y_i \text{ for every assigned training context } i\}. \tag{5}$$

This uses only the eight distinct training contexts and their oracle labels. An independent double-precision filter agrees with the original oracle on all $4^6 = 4,096$ sequences. All 768 entity/arm assignments across Q3/Q1/Q4/FAM3R2 match the semantic streams and metadata, with 48 application presentations per name. Archived compiled-stream receipts provide an input provenance link, not a replay of optimizer reads. The original review script (ID0) was unavailable; these counts are an independent reconstruction of the stated audit, not a claimed rerun of that missing artifact.

CROSS8 leaves a unique endpoint for every trained name. SAME8 leaves one for 73/96, 85/96, 83/96 and 83/96 names in Q3, Q1, Q4 and FAM3R2, respectively; every other name has two candidates. Q3 is shared by the Llama and first Phi block. FAM3R2 is a new assignment in the same factual world. Original history clusters are recovered through the preserved pair mapping; they are not inferred by halving selected name counts.

Table 8: Endpoint identification from application labels alone (post hoc design audit). Each arm has 96 application-trained names. Joint-unique names retain all 48 original history clusters in every bank; whole-pair counts are reconstructed from member IDs. Q3 is shared by Qwen W3, Llama, and Phi 1. FAM3R2 is Phi 2's assignment in the same world.

| Bank | S unique | C unique | S ambiguous | Whole joint pairs |
|---|---|---|---|---|
| Q3 | 73 | 96 | 23 | 25 |
| Q1 | 85 | 96 | 11 | 37 |
| Q4 | 83 | 96 | 13 | 35 |
| FAM3R2 | 83 | 96 | 13 | 35 |

**Membership and estimands.** JOINT_UNIQUE requires the true endpoint to be the unique candidate in both arms. S_AMBIGUOUS_C_UNIQUE requires two SAME8 candidates and one CROSS8 candidate. OTHER retains every unexpected configuration and is empty in these data. A whole-pair sensitivity retains only original pairs whose two members are JOINT_UNIQUE. Full K/S/C joins cover 768 ordinary DIRECT test cases per block, without validity or correct-recall filtering. All 18 final prediction files were scanned and all original primary accuracies reproduced. The companion outputs retain each row's identity, original cluster, context, endpoint and arm outcomes.

Within each view, the primary post hoc estimand first averages within each represented original history cluster and then averages cluster means equally. Consequently, an unweighted net-case count divided by the subset size need not equal this estimand when selection splits pairs. Entity-balanced means are reported separately. The standalone pair bootstrap uses 2,000 draws and seed 61991, with shared draws across arms. The new crossed sensitivity uses shared Poisson(1) weights for original history clusters and visible contexts (seed 2026092401). Context-weighted means are normalized within each included cluster before history weights are applied, preserving equal-cluster weighting. Zero-context-mass clusters receive zero effective weight; shared draws with zero total weight in any required nonempty view are redrawn. This explicitly new post hoc estimator does not replace the archived full-panel intervals.

The history/context product weights are related to crossed-factor bootstrap methods (Owen & Eckles, 2012). Our within-history normalization and subset selection are additional choices. The reference does not establish their finite-sample percentile coverage, and the conditional interpretation remains unchanged.

Direct difference-of-effects contrasts are predefined: FULL minus JOINT_UNIQUE, FULL minus whole-pair JOINT_UNIQUE, and ambiguous minus JOINT_UNIQUE, for $C - S$ and $C - K$. Pair intervals for these differences use shared multinomial multiplicities on the full 48-cluster frame (2,000 accepted draws, seed 2026092402), renormalized within each view. Crossed contrasts share the accepted crossed draws. These differences, rather than differences in significance status, assess attenuation or heterogeneity. Empty strata remain explicitly unevaluable. No architecture pooling, mediated fraction or outcome-selected threshold is used.

**Interpretation.** JOINT_UNIQUE and the complete-pair sensitivity retain positive $C - S$ effects under both uncertainty methods in every block. The full-minus-JOINT point contrasts are negative rather than attenuated; their uncertainty is retained with the complete strata. The narrow conclusion is that positive learning effects are not confined to names with endpoint ambiguity under SAME8 application labels. The groups remain selected by realized training designs; the result does not rule out relearning, unequal information, optimization difficulty or entity-specific computation. Optional answer-seen/answer-unseen counts are separately reported as descriptive, treatment-dependent supports, not new randomized comparisons.

The supplement includes the frozen protocol, memberships, all six joined observation banks and full estimates. J's absolute K/S/C levels use the same history-cluster weighting and are in its estimates directory. The README provides commands for aggregate display and recorded-observation reproduction. No new inference is needed.

Table 9: All post hoc label-identifiability strata on ordinary DIRECT test cases. F: full panel; J: both curricula uniquely identify the endpoint; A: SAME8 ambiguous/CROSS8 unique; W: whole original pairs with both names in J; O: other configurations. E/H gives entities/history clusters. Effects are history-cluster-balanced CROSS8−SAME8 accuracy differences in percentage points. Intervals are the new pair bootstrap and explicitly new cluster-balanced crossed Poisson sensitivity, not replacements for archived headline intervals. Empty O strata are retained. Exact K/S/C levels, C−K contrasts, gains/breaks and entity-balanced sensitivities are supplied in `identifiability/estimates/identifiability_outcomes.csv`. Memberships and joined observations are indexed in the supplement README.

| Block | Stratum | E/H | Cases | C−S | Pair 95% | Crossed 95% |
|---|---|---|---|---|---|---|
| Qwen W3 | F | 96/48 | 768 | 41.54 | [37.24, 45.96] | [36.06, 47.46] |
| Qwen W3 | J | 73/48 | 584 | 47.92 | [42.06, 54.04] | [39.62, 55.76] |
| Qwen W3 | A | 23/23 | 184 | 25.54 | [15.22, 35.33] | [11.61, 40.89] |
| Qwen W3 | W | 50/25 | 400 | 44.00 | [37.25, 50.75] | [36.35, 51.97] |
| Qwen W3 | O | 0/0 | 0 | — | — | — |
| Qwen W1 | F | 96/48 | 768 | 31.12 | [26.82, 35.16] | [25.51, 36.57] |
| Qwen W1 | J | 85/48 | 680 | 32.16 | [27.73, 36.46] | [25.56, 38.77] |
| Qwen W1 | A | 11/11 | 88 | 27.27 | [17.05, 37.50] | [9.02, 47.44] |
| Qwen W1 | W | 74/37 | 592 | 30.91 | [25.51, 36.15] | [24.06, 37.35] |
| Qwen W1 | O | 0/0 | 0 | — | — | — |
| Qwen W4 | F | 96/48 | 768 | 37.89 | [34.24, 41.41] | [32.29, 42.68] |
| Qwen W4 | J | 83/48 | 664 | 40.89 | [36.85, 44.79] | [34.45, 47.03] |
| Qwen W4 | A | 13/13 | 104 | 25.00 | [13.46, 34.62] | [7.42, 43.49] |
| Qwen W4 | W | 70/35 | 560 | 38.57 | [34.64, 42.32] | [32.14, 44.00] |
| Qwen W4 | O | 0/0 | 0 | — | — | — |
| Llama W3 | F | 96/48 | 768 | 38.41 | [33.46, 43.36] | [32.43, 44.68] |
| Llama W3 | J | 73/48 | 584 | 45.44 | [39.97, 50.91] | [37.99, 53.45] |
| Llama W3 | A | 23/23 | 184 | 16.30 | [6.52, 25.54] | [1.77, 31.04] |
| Llama W3 | W | 50/25 | 400 | 45.25 | [40.00, 50.25] | [37.26, 52.78] |
| Llama W3 | O | 0/0 | 0 | — | — | — |
| Phi R1 | F | 96/48 | 768 | 38.54 | [35.03, 42.06] | [33.09, 43.13] |
| Phi R1 | J | 73/48 | 584 | 46.22 | [41.93, 50.39] | [38.78, 53.47] |
| Phi R1 | A | 23/23 | 184 | 17.93 | [9.24, 26.09] | [4.89, 31.76] |
| Phi R1 | W | 50/25 | 400 | 42.75 | [38.75, 46.25] | [36.51, 47.84] |
| Phi R1 | O | 0/0 | 0 | — | — | — |
| Phi R2 | F | 96/48 | 768 | 41.15 | [37.50, 44.53] | [36.14, 45.61] |
| Phi R2 | J | 83/48 | 664 | 43.49 | [39.84, 47.01] | [37.66, 49.56] |
| Phi R2 | A | 13/13 | 104 | 27.88 | [17.31, 38.46] | [11.87, 44.90] |
| Phi R2 | W | 70/35 | 560 | 42.86 | [39.10, 46.96] | [37.32, 48.04] |
| Phi R2 | O | 0/0 | 0 | — | — | — |

Table 10: Predefined direct differences between CROSS8−SAME8 effects (percentage points), using common resampling weights. F−J and F−W are full minus restricted effects; positive values would indicate attenuation in the restriction. A−J is stratum heterogeneity. These are post hoc population contrasts, not randomized mediation. Pair intervals use a shared full-history sampling frame; crossed intervals use the cluster-balanced extension. Different significance statuses alone are not interpreted as different effects.

| Block | Contrast | Difference | Pair 95% | Crossed 95% |
|---|---|---|---|---|
| Qwen W3 | F−J | -6.38 | [-10.68, -2.60] | [-12.43, -0.09] |
| Qwen W3 | F−W | -2.46 | [-6.96, 1.93] | [-8.03, 2.51] |
| Qwen W3 | A−J | -22.37 | [-34.77, -9.90] | [-40.77, -3.09] |
| Qwen W1 | F−J | -1.04 | [-2.86, 0.52] | [-4.76, 2.62] |
| Qwen W1 | F−W | 0.21 | [-1.49, 1.93] | [-2.98, 3.29] |
| Qwen W1 | A−J | -4.89 | [-16.20, 7.42] | [-24.98, 18.35] |
| Qwen W4 | F−J | -2.99 | [-5.34, -1.04] | [-7.15, 0.86] |
| Qwen W4 | F−W | -0.68 | [-3.19, 1.69] | [-4.30, 2.37] |
| Qwen W4 | A−J | -15.89 | [-27.61, -5.21] | [-35.25, 3.98] |
| Llama W3 | F−J | -7.03 | [-10.94, -3.39] | [-12.70, -1.92] |
| Llama W3 | F−W | -6.84 | [-11.63, -2.19] | [-13.37, -0.38] |
| Llama W3 | A−J | -29.14 | [-40.33, -18.36] | [-45.76, -12.05] |
| Phi R1 | F−J | -7.68 | [-11.46, -4.43] | [-13.22, -2.12] |
| Phi R1 | F−W | -4.21 | [-7.96, -0.89] | [-9.10, 0.37] |
| Phi R1 | A−J | -28.29 | [-38.05, -19.29] | [-44.66, -10.85] |
| Phi R2 | F−J | -2.34 | [-4.56, -0.39] | [-6.13, 1.30] |
| Phi R2 | F−W | -1.71 | [-3.97, 0.36] | [-4.97, 1.43] |
| Phi R2 | A−J | -15.60 | [-26.43, -3.65] | [-33.44, 3.97] |

Table 11: Forward function replacement on intervention-held-out trained names. The transfer panel uses 48 names × eight contexts (384 cases; BF16) with crossed intervals. The explanatory panel uses 48 names × sixteen contexts (768 cases; FP32) with history-pair intervals. Gains are pp; these are distinct generations and supports.

| Block | Transfer panel: gain [95%] | Explanatory panel: gain [95%] |
|---|---|---|
| Qwen W3 | 10.94 [5.86,16.68] | 9.38 [6.35,12.67] |
| Qwen W1 | 1.56 [0.30,4.24] | 1.69 [0.91,2.74] |
| Qwen W4 | 3.91 [1.07,8.42] | 5.47 [3.21,8.35] |
| Llama W3 | 1.82 [0.30,4.90] | 0.78 [−0.33,2.58] |

## G SELECTION, INTERVENTIONS AND EXPLANATORY CONTROLS

The corrected selection considers six depth windows, 0–5, 6–11, 12–17, 18–23, 24–29 and 30–35, for both attention and MLP modules in each direction. On Q3 SELECT, forward composition eligibility requires a 5-point DIRECT gain and either a 0.03 alias-TV reduction or a 5-point fact-sensitive both-correct gain, with recall/explicit losses at most 3 points and validity loss at most 1 point. Reciprocal name-influence screening uses its signed loss and a 0.05 TV increase. Eligible candidates are ordered by exact effect, reciprocal support, parameter count, layer count, starting layer and ID. Only forward MLPs 12–17 qualify. All six individual children are measured; SELECT advances layers 17, 12 and 13 alongside the group. VERIFY retains the group: no lower-parameter child is within the strict 2-point substitution tolerance. There is no reverse-track nominee. Q1/Q4 retain the location; Llama receives the declared interval mapping $[\lfloor 12 \cdot 28/36 \rfloor, \lceil 18 \cdot 28/36 \rceil) = [9, 14)$. The accepted replay is retrospective and retains prior TEST exposure.

Function replacement evaluates each donor MLP on its recipient's live normalized input. Complete internal MLP maps travel together; normalization outside the MLP, attention, embeddings and the output head remain native. Function-transfer evaluations use BF16 weights with FP32 reductions; explanatory interventions use their separately qualified FP32 path. Explanatory masks use each prompt's actual final non-padding index: PRED selects that position, PREFIX the remaining real positions, and ALL their union. The 384-case function/reader panel has 48 application-trained names and eight contexts, whereas the 768-case explanatory panel has the same number of names and 16 contexts. Neither constitutes a new training replicate.

Reader hooks replace MLP outputs or one head's slice before `o_proj`; the shared projection bias is applied once. Clean references are retained in FP32, clamped at all non-padding positions and cast to the live output dtype. Every condition recomputes the full sequence without a KV cache. Self-clamp and restoration checks accompany three norm-matched Gaussian controls (seeds 11, 23 and 37); zero changes remain zero. Correctness and target probability interactions are reported separately for all 16 readers.

For the additive approximation, define $r(x)$ as the highest-scoring incorrect clean label, with token-ID tie breaking, and $m_X(x) = \ell_{X,y}(x) - \ell_{X,r(x)}(x)$. Figure 5B uses

$$\rho_{\mathrm{rival}} = \sqrt{\frac{\mathbb{E}_x[(m_G - m_{\mathrm{add}})^2]}{\mathbb{E}_x[(m_G - m_0)^2]}}. \tag{6}$$

The rival is fixed from CLEAN. Compatibility requires $\rho_{\mathrm{rival}} \leq 0.2$, centered residual-RMSE ratio at most $0.5$, legal-label agreement at least $0.95$, and both observed shifts above qualified numerical floors. Q4/C satisfies every criterion; Q3/C fails the rival criterion despite high legal-label agreement. Additivity uses actual input-specific single-layer outputs and does not specify outside-label logits.

Fixed output comparators fit four intercepts with an optional SAME8-anchor contrast on CAL only, using FP64 weighted-mean ridge with coefficient penalty $0.001$. They receive no current oracle answer or endpoint. Outside logits stay fixed. Changed-case fidelity is exact unrestricted token agreement among hybrid-changed cases (the declared minimum is $80\%$); accuracy and validity tests retain the full panel. These denominators and scales differ from additive legal-label agreement, so their raw agreement percentages are not interchangeable.

Table 12: Direct paired accuracy controls on intervention-held-out application-trained names. $S$, $C$, and $K$ denote SAME8, CROSS8, and their shared parent. Donor is $\mathrm{Acc}(S[C_G]) - \mathrm{Acc}(S[K_G])$. Location is $\mathrm{Acc}(S[C_G]) - \mathrm{Acc}(S[C_{G_{\mathrm{hash}}}])$ in A and substitutes $G_{\mathrm{neighbor}}$ in B. Reciprocal is $\mathrm{Acc}(C[S_G]) - \mathrm{Acc}(C)$. Net is the signed additional correct-case count. Differences and intervals are percentage points. These are copied paired contrasts, not differences between interval endpoints.

**A. Function-transfer TEST: 384 cases, 48 names, 24 history pairs, 8 contexts.**

| Block | Contrast | Net | Difference | Crossed 95% |
|---|---|---|---|---|
| Qwen W3 | Donor | $+38$ | $+9.90$ | $[4.27, 16.10]$ |
| Qwen W3 | Location | $-6$ | $-1.56$ | $[-6.85, 3.51]$ |
| Qwen W3 | Reciprocal | $-4$ | $-1.04$ | $[-3.06, -0.13]$ |
| Qwen W1 | Donor | $+1$ | $+0.26$ | $[-2.99, 3.42]$ |
| Qwen W1 | Location | $-6$ | $-1.56$ | $[-4.62, 0.48]$ |
| Qwen W1 | Reciprocal | $-4$ | $-1.04$ | $[-3.36, -0.09]$ |
| Qwen W4 | Donor | $+0$ | $+0.00$ | $[-0.98, 1.05]$ |
| Qwen W4 | Location | $-29$ | $-7.55$ | $[-12.83, -3.56]$ |
| Qwen W4 | Reciprocal | $+3$ | $+0.78$ | $[0.06, 2.79]$ |
| Llama W3 | Donor | $+3$ | $+0.78$ | $[0.04, 2.94]$ |
| Llama W3 | Location | $-14$ | $-3.65$ | $[-8.07, -1.04]$ |
| Llama W3 | Reciprocal | $-1^*$ | $-0.26$ | $[-1.3762, -0.0015]$ |

**B. Explanatory TEST: 768 cases, 48 names, 24 pairs, 16 contexts.**

| Block | Contrast | Net | Difference | History 95% | Crossed 95% |
|---|---|---|---|---|---|
| Qwen W3 | Donor | $+62$ | $+8.07$ | $[5.32, 10.91]$ | $[4.70, 12.46]$ |
| Qwen W3 | Location | $+33$ | $+4.30$ | $[2.51, 6.39]$ | $[1.87, 7.74]$ |
| Qwen W1 | Donor | $+3$ | $+0.39$ | $[-0.72, 1.31]$ | $[-1.77, 2.21]$ |
| Qwen W1 | Location | $-13$ | $-1.69$ | $[-3.01, -0.63]$ | $[-3.94, -0.27]$ |
| Qwen W4 | Donor | $+1$ | $+0.13$ | $[-0.68, 0.98]$ | $[-1.27, 1.63]$ |
| Qwen W4 | Location | $-35$ | $-4.56$ | $[-6.89, -2.59]$ | $[-7.99, -2.07]$ |
| Llama W3 | Donor | $-4$ | $-0.52$ | $[-1.37, -0.07]$ | $[-1.82, -0.05]$ |
| Llama W3 | Location | $-7$ | $-0.91$ | $[-2.60, -0.01]$ | $[-2.96, 0.21]$ |

The selected $G$ spans Qwen layers 12–17 and Llama layers 9–13. The hash alternative $G_{\mathrm{hash}}$ spans Qwen 0–5 and Llama 0–4. The neighbor $G_{\mathrm{neighbor}}$ spans Qwen 18–23 and Llama 14–18. Layer indices are zero-based and inclusive. A uses the recorded crossed history/context primary interval. B reports primary history and crossed sensitivity separately. $^*$Llama reciprocal has zero gains and one break, with canonical sparse-outcome status. The extra decimal places preserve its near-zero bound. Its retained legacy Poisson sensitivity includes zero. This does not support a strong population or reciprocal-equivalence claim.

Table 13: Fixed output comparators on the explanatory TEST panel: 48 application-trained names $\times$ 16 contexts (768 cases), with CROSS8 (C) or parent K donors. Fidelity uses only hybrid-changed cases and exact unrestricted token identity; accuracy differences use all 768 cases and are in percentage points. Brackets are recorded history-cluster 95% intervals. Every displayed comparator fails its full acceptance contract.

| Block | Donor | Model | Changed fidelity | Acc.−hybrid [95%] |
|---|---|---|---|---|
| Qwen W3 | C | Global | 2/77 | $-9.11\,[-12.34, -6.13]$ |
| Qwen W3 | C | Global+anchor | 20/77 | $-6.77\,[-9.62, -4.27]$ |
| Qwen W3 | K | Global | 2/21 | $-1.17\,[-2.81, 0.45]$ |
| Qwen W3 | K | Global+anchor | 4/21 | $-0.13\,[-1.74, 1.57]$ |
| Qwen W1 | C | Global | 0/17 | $-1.69\,[-2.73, -0.91]$ |
| Qwen W1 | C | Global+anchor | 4/17 | $-1.30\,[-2.28, -0.63]$ |
| Qwen W1 | K | Global | 1/20 | $-1.30\,[-2.70, -0.19]$ |
| Qwen W1 | K | Global+anchor | 8/20 | $-0.39\,[-1.59, 0.52]$ |
| Qwen W4 | C | Global | 5/49 | $-5.21\,[-8.00, -3.13]$ |
| Qwen W4 | C | Global+anchor | 15/49 | $-4.04\,[-6.50, -2.26]$ |
| Qwen W4 | K | Global | 3/61 | $-5.08\,[-7.62, -3.14]$ |
| Qwen W4 | K | Global+anchor | 27/61 | $-2.08\,[-4.17, -0.73]$ |
| Llama W3 | C | Global | 1/10 | $-0.65\,[-2.47, 0.43]$ |
| Llama W3 | C | Global+anchor | 4/10 | $-0.26\,[-1.74, 0.63]$ |
| Llama W3 | K | Global | 2/16 | $-1.43\,[-3.40, -0.32]$ |
| Llama W3 | K | Global+anchor | 7/16 | $-0.52\,[-1.68, 0.17]$ |

Table 14: Input-dependent additive score approximation on the same 768-case explanatory TEST panel. Entries are history-cluster 95% intervals. Compatibility requires residual-RMSE ratio $\leq 0.5$, clean-rival error ratio $\leq 0.2$, legal-label agreement $\geq 0.95$, and both measured shifts above qualified numerical floors. Only Qwen W4/C satisfies every criterion. This legal-score approximation does not model outside-label logits.

| Block | Donor | RMSE ratio | Rival ratio | Legal agreement |
|---|---|---|---|---|
| Qwen W3 | C | $[0.329, 0.397]$ | $[0.283, 0.340]$ | $[0.969, 0.990]$ |
| Qwen W3 | K | $[0.223, 0.282]$ | $[0.208, 0.295]$ | $[0.982, 0.998]$ |
| Qwen W4 | C | $[0.157, 0.189]$ | $[0.124, 0.157]$ | $[0.996, 1.000]$ |
| Qwen W4 | K | $[0.231, 0.269]$ | $[0.206, 0.271]$ | $[0.975, 0.994]$ |

**Fixed-shift feature coding and fitting target.** For case $i$ and legal label $j$, the recorded anchor feature is $a_{ij} = \mathbf{1}[j = a_i] - 1/4$ when a SAME8 training anchor $a_i$ exists. Its entire row is zero for names without an anchor. GLOBAL4 predicts $b_j$ and GLOBAL4_PLUS_ANCHOR predicts $b_j + \gamma a_{ij}$, with one shared anchor coefficient. The calibration target is the hybrid-minus-clean change in label-to-outside log odds $r_{ij} = \ell_{ij} - \mathrm{LSE}(\ell_{i,\mathrm{outside}})$, rather than the centered logits used by the additive approximation. The FP64 fit minimizes the weighted mean squared error over cases and four labels plus $0.001\|\beta\|_2^2$ for every fitted coefficient. Calibration assigns equal total weight to treated-development and fact-only-validation populations, equal weight to original histories within each population and equal weight to cases within each history. The fitted shift is added to clean legal logits while outside-token logits remain unchanged. Coefficients are fixed for verification and test. These are recorded definitions and no new fit is performed for this presentation.

### G.1 EXPLANATORY-INTERVENTION DETAILS AND PRESERVED DISPLAY

**Function replacement and controls.** We test consequences of replacing learned transformations at inference. A donor MLP evaluates $f_\ell^C(z_\ell^S)$ in place of $f_\ell^S(z_\ell^S)$ on the recipient's live normalized input, which incorporates earlier replacements. Complete MLP maps and internal biases/buffers travel together; external normalization and all other components remain recipient-native. No donor question or oracle value is supplied. This differs from relocating cached states through back- or self-patching (Biran et al., 2024; Dai et al., 2026). Qwen MLPs 12–17 are selected on W3 development support and frozen for W1/W4; Llama uses MLPs 9–13. Selection history, inventories and hooks are in Appendix G.

The function-transfer panel has 384 cases from 48 intervention-held-out trained names and eight contexts. The explanatory panel has 768 cases from the same number of names and sixteen contexts. Their separate BF16/FP32 execution contracts do not make them interchangeable replicas. Forward replacements produce partial gains on both panels, with the Llama explanatory-panel effect unresolved (Table 11). Qwen's function-transfer world-mean interval is $[-6.65, 17.59]$ pp.

The controls qualify donor and location specificity (Table 12). On the explanatory panel, the CROSS8 donor exceeds the parent donor by 8.07 pp in W3, with crossed interval $[4.70, 12.46]$, but this advantage is not consistent across blocks. Alternative locations sometimes perform better than the selected group. Reciprocal effects are heterogeneous and can depend on very few changed answers. These results identify effective interventions without establishing a uniquely privileged location for the learned behavior.

**Prefix versus prediction position.** CLEAN, prediction-only, prefix-only and all-position masks use each prompt's actual prediction index; prefix excludes that index and padding. Qwen prefix replacement produces nearly all the observed net accuracy gain: W3 gains 69 cases versus 72 with all positions, W1 12 versus 13, and W4 41 versus 42. Prediction-only gains are 2/0/3 (Figure 5). Closeness is not equivalence, and prefix contribution does not itself identify retrieval. The factorial interaction ALL$-$PRED$-$PREFIX$+$CLEAN can reflect top-1 thresholds even when score changes are additive.

**Additive scores and fixed output shifts.** The additive approximation uses centered four-label logits from clean and isolated-layer replacements, without fitted coefficients:

$$\widehat{\ell}_G(x) = \ell_0(x) + \sum_{j \in G} \big(\ell_j(x) - \ell_0(x)\big). \tag{7}$$

The highest-scoring incorrect label in the clean model defines a fixed rival, reused for all hybrid and additive predictions. Qwen W4 with a CROSS8 donor meets the declared score-fidelity criteria; W3 does not. Legal-answer agreement alone is insufficient for score fidelity, and a failed tolerance does not imply zero explanatory value. This predictor is input-dependent and supplies no unrestricted outside-token model.

GLOBAL4 fits four constant shifts to the answer-label scores. GLOBAL4_PLUS_ANCHOR adds one coefficient tied to the answer associated with that name in SAME8 training. Both fit the hybrid's score changes on calibration examples without receiving the current target or missing endpoint. Outside-token logits remain unchanged. Appendix G gives the recorded feature coding, weighting and ridge penalty.

On treated-held-out TEST, the anchor comparator matches only 20/77 hybrid-changed cases in W3 and 15/49 in W4. Their failure excludes these comparator forms, not all output-level explanations or alternative algorithms.

**Downstream dependence.** With $Y_{00}$ clean, $Y_{10}$ replacement, $Y_{01}$ clean self-clamp and $Y_{11}$ replacement plus clamp, the reader interaction is $(Y_{10} - Y_{00}) - (Y_{11} - Y_{01})$. Restoring a clean reader output tests dependence within this intervention regime; the self-clamp checks execution identity. W3 head 35:25 yields accuracy/probability interactions of 2.08/2.50 pp and W4 MLP35 yields 2.60/2.03 pp, separated from three matched controls. Other readers are heterogeneous, including probability-only and sparse one-case Llama effects (Table 7). Causal-abstraction accounts require testing an alignment between low- and high-level models (Geiger et al., 2025). Subspace-patching counterexamples (Makelov et al., 2024) also caution against inferring natural necessity

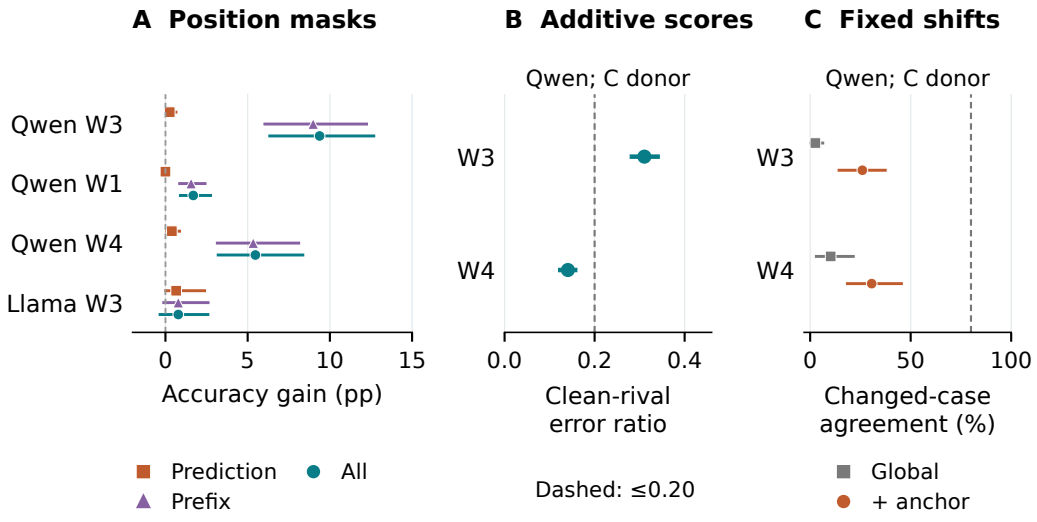



Figure 5: Explanatory interventions on 768 cases per comparison: 48 intervention-held-out trained names with sixteen contexts. A: accuracy gains over the unmodified model (CLEAN) when replacements affect different input positions. B: additive approximation error, measured by root mean square (RMS) clean-rival error relative to the observed joint score shift (criterion $\leq 0.20$). C: fitted fixed-shift exact-token agreement on cases where the hybrid changed the answer (criterion $\geq 80\%$). "C donor" denotes imported CROSS8 (4:4) functions. "Global" fits four constant answer-score shifts; "+ anchor" also fits a coefficient for the name's SAME8 training answer. Intervals are recorded history-cluster 95%. Outcomes and denominators differ across panels; these thresholds are only part of the full fidelity criteria.

from intervention success. These works motivate our qualifications without diagnosing our whole-module swaps as illusory. Native-state changes are mostly sparse; no transcoder feature claim is made. Corrected analyses retain prior TEST inspection and do not become prospective mechanism tests.

# H GRADED ASSIGNMENT AND COMMON ANSWER-UNSEEN EXTENSION

This subsequent extension trained six Qwen children and reduced existing observations for the original six behavioral blocks. It did not retrain the historical SAME8/CROSS8 conditions, add a model family, or run new component interventions. We distinguish completed execution from compliance with the original numerical requirements. All planned M1/M2 training and both full analyses completed, but the original native-output and batch-invariance gates remained failed. The integration basis is an explicitly amended measurement protocol, not an unchanged prospective pass. The optional Phi conflict-error decomposition was disabled. The Qwen W4 supplied-conflict cost is part of the required graded evaluation and is not that optional module.

Design and execution are in Section H.1, numerical qualifications in Section H.2, supports and estimators in Section H.3, complete results in Section H.4, and release/reproduction in Section H.5.

## H.1 FIXED CONSTRUCTION, DOSE AND PROVENANCE

The extension uses the same Qwen W3/W1/W4 K parents and original application plans. Let $m \in \{0, 1, 2, 4\}$ be the number of exchanged context pairs among the four historically exchanged positions, indexed 4–7. Their order is fixed by the hexadecimal SHA-256 digest of the UTF-8 string formed by joining the protocol salt, canonical plan digest, pair ID and index with the separator `|`; the index breaks a tie. Canonical plan serialization sorts keys, uses compact JSON separators and UTF-8. M1 takes the first position, M2 the first two, and M4 all four. The same masks apply across cycles and renderings, without consuming the trainer's random stream. The included reference implementation and masks specify these operations exactly; they were not optimized for outcomes or endpoint identifiability.

Each new semantic stream starts from the original SAME8 step sequence and changes only the two application rows' entity assignments at masked positions. Contexts, oracle targets, auxiliaries and within-step semantic material are otherwise preserved. Every new child begins from its own world's K checkpoint with a fresh optimizer. The fixed dose is 2,304 updates and 18,432 examples: 4,608 application, 9,216 generic explicit and 4,608 factual. Each name receives eight distinct application contexts and 48 presentations, so M1 has six minority presentations and M2 twelve. The inherited BF16/AdamW8bit recipe uses learning rate $2 \times 10^{-5}$, 50 warmup updates followed by a constant schedule, zero weight decay and gradient-norm clipping at 1.0. Microbatch four with accumulation two gives effective batch eight. Within-world seeds are 61003/61001/61004 for W3/W1/W4. New levels are not independent optimizer replications or sequentially trained stages.

The fixed order was W3 M1, W1 M2, W4 M1, W3 M2, W1 M1, W4 M2. Execution records preserve actual attempts, compiled-stream identities, per-step consumed prefixes and final checkpoint hashes. Independent training checks validated 13,824 update-prefix hashes and 110,592 finite per-example losses across the six runs. Charged GPU time, including repeated and failed work, was 4.4025 hours; operator active time and process CPU time were not directly measured. Native tokenization and input audits distinguish semantic matching from identical padding or compute. The final checkpoint is used without outcome-based selection or extra training.

## H.2 NUMERICAL FAILURES AND THE AMENDED MEASUREMENT POLICY

The original BF16 qualification failed cross-batch probability and native-output requirements (original acceptance checks A14 and A15). Amendment A relaxed continuous cross-batch gating while retaining discrete checks. Following a discrete failure, B introduced a 99.5% criterion. Amendment C removed new-child native/regroup checks from gating and required complete singleton sensitivity. These changes followed observed qualification failures and do not constitute a predeclared successful qualification. The internal execution record retains the failure/amendment chronology and distinguishes reported author-applied changes from implementation-side interpretations of authorization.

The recorded primary path and the uniform singleton path are both retained. Historical incumbent predictions used a larger original panel than the new children, so an identical batching algorithm did not imply identical batch composition. Singleton rescoring covers K, M0, M1, M2 and M4 in all three Qwen worlds. Cached Llama/Phi observations are unchanged between these two anal-

yses; their appearance in both directories does not add measurements or replications. Historical manuscript 8:0/4:4 values and intervals are preserved rather than silently replaced by singleton results.

Sampled same-batch repeatability and the bounded native/singleton witness checks are separate from full cross-batch invariance. The witness check does not establish native-generation agreement on every row or equality of continuous probabilities. Original batched-native disagreements for W3 M2 and W4 M2 remain recorded. The large M1–M0 effect survives both paths, but secondary classifications can change. For example, W1 M1–M4 has crossed bounds $[-5.1643, 1.0045]$ pp on the recorded path and $[-4.9717, 1.0971]$ under singleton measurement. Rounding both lower bounds to $-5.0$ would obscure a different descriptive compatibility decision.

### H.3 Cases, restrictions and uncertainty

The primary panel uses the original 768 trained-name DIRECT TEST cases: 96 names and 48 original history clusters. It is separate from the 768-case intervention panel. All fixed cases are retained, with invalid full-vocabulary outputs counted as errors. Secondary ordinary cells retain their original population, partition and rendering, alongside recall, alias grids and endpoint-balanced lattices. Each new-child scoring bank additionally includes 768 final-checkpoint application-training diagnostics; these did not select checkpoints or extend training. M1's minority subset contains 96 distinct cases and M2's 192, rather than their repeated presentations.

For name $e$, let $\mathcal{A}_e^0$ and $\mathcal{A}_e^4$ be its original M0 and M4 application-training labels. The common ANSWER-UNSEEN view retains a fixed test case if its oracle target is outside $\mathcal{A}_e^0 \cup \mathcal{A}_e^4$. Membership therefore uses training labels and fixed test targets, not model predictions. It is post hoc and contains only P/S endpoints on already inspected support. Names received application training and labels are familiar globally. Its internal name `U` must not be confused with the older additional fact-only population. Complete memberships, endpoint/target counts and original history identities are supplied.

The all-level identifiable restriction retains names whose true endpoint is uniquely implied by the application labels at every application level M0/M1/M2/M4. A whole-pair sensitivity requires both members of the original history pair. These are new restrictions, separate from the original M0/M4 identifiability tables. Another supplementary view intersects common answer-unseen with the original joint-unique restriction; its sparse support must not replace the full answer-unseen view. Empty views remain explicitly unevaluable.

For ordinary accuracy and alias summaries, the point estimate averages within represented original histories and then weights those history means equally. On unequal subsets, raw correct count divided by row count need not equal this estimate. Standalone history intervals use multinomial resampling of represented histories, shared across levels. The crossed sensitivity applies independent Poisson(1) weights to histories and visible contexts, normalizes context-weighted outcomes within each history, and then averages with effective history weights. A history with zero context mass receives zero effective weight. Draws with zero total mass in any required nonempty view are rejected jointly. Difference-of-effect sensitivities use a shared full-history frame rather than subtracting independently resampled view intervals. There are 10,000 accepted draws with seed 2026092508; sorted identities and blockwise random-number layout are fixed in the producer.

The three worlds by three adjacent contrasts (M1–M0, M2–M1 and M4–M2) form the nine-comparison family. Bonferroni sensitivities use percentile confidence $1 - 0.05/9$ for each contrast under both schemes. This is an approximate bootstrap multiplicity sensitivity, not independent validation of finite-sample coverage. Secondary, parent-relative and compatibility comparisons retain their stated unadjusted intervals. Descriptive similarity to M4 requires both paired 95% intervals to lie within $\pm 5$ pp; the complete component criterion instead uses the inherited $\pm 3$ pp margin in every required cell. Failure of such a two-sided criterion can reflect loss, imprecision or improvement beyond the margin. It is neither blanket harm nor evidence of equivalence.

The separate equal-world summary weights the three fitted-world effects equally and uses a $t$ sensitivity with two degrees of freedom. It does not pool cases across worlds or measure optimizer-seed variability. Lattice points and their context-primary intervals retain the original endpoint-balanced weighting and fixed $B_0$; their crossed sensitivity uses the inherited history/context row-product

weights. No lattice ceiling is transferred to the answer-unseen subset. These definitions are new reduction provenance and do not replace previously reported original-comparison intervals.

### H.4 Complete results and interpretation

All planned levels and both Qwen paths are represented in the tables and full-precision supplement. M1–M0 gains are 35.68/29.17/35.42 pp on the recorded path. Their adjusted crossed intervals are $[26.93, 44.34]$, $[20.91, 36.60]$ and $[27.80, 41.94]$. The equal-world gain is 33.42 pp with separate $t$ interval $[24.26, 42.58]$; singleton measurement gives 33.46 pp $[24.21, 42.72]$. M1 improves above K on the full trained-name support. These are benefits at a tested assignment, not a universal minimum or a fitted transition.

M2–M1 is 4.95/0.00/4.04 pp on the recorded path. W3/W4 ordinary intervals are positive, but their adjusted crossed intervals and the separate world-level interval include zero. M4–M2 is unresolved throughout; W4's M2 point exceeds M4, so monotonicity or optimality is not established. Only W3 M2 satisfies the descriptive five-point criterion on both paths and schemes. W3 M1 remains 5.86/5.60 pp below M4 on the recorded/singleton paths. Training-minority fits are high, but fitting a minority context is not itself proof of generalization.

On full answer-unseen support, M4–M0 is positive under both ordinary interval schemes in five of six original blocks, with Phi R1 unresolved. No block establishes positive M4–K under both schemes; both Phi parent contrasts are negative. For graded Qwen, W1 M1–M0 has crossed lower bound exactly zero, and W1 M2–K is negative under both schemes and paths. The more favorable sparse intersection cannot substitute for these results. All-level identifiable and complete-pair restrictions retain positive M1–M0 effects, constraining endpoint ambiguity as a sole explanation without proving mediation or equal learning information.

Original fact-only TEST M1 accuracies are 48.31/57.36/73.89%, with parent-relative gains of 18.42/11.91/36.26 pp; both schemes and paths retain these secondary positive effects. Additional fact-only names are separate, with W3 M1–K unresolved. W4 supplied-conflict accuracy drops to 80.08% at M1 and 86.46% at M2 from M0's 95.31%. This adverse effect survives singleton scoring; it is not a necessary trade-off across worlds or evidence that an internal memory source was selected. All six new children have 100% observed primary factual recall while comprehensive component compatibility remains unestablished.

**Reading the graded-assignment tables.** B denotes the recorded analysis path. S denotes uniform singleton scoring of the fifteen Qwen conditions. Llama and Phi retain the same cached observations in both analysis exports. History and crossed columns are conditional intervals for fixed checkpoints. Accuracy is the equal-history mean unless the caption states otherwise. $n/h/c$ gives cases, original history clusters, and distinct contexts. All invalid outputs remain errors. Original numerical gates A14/A15 remain failed. These tables describe the qualified amended-protocol generation, not restored batch invariance.

Table 15: Matched graded design and label-only endpoint identifiability.

| World | Allocation | Training | Names | Minority | Unique | Updates |
|---|---|---|---|---|---|---|
| Qwen W3 | 8:0 | Historical | 96 | 0 | 73 | 2304 |
| Qwen W3 | 7:1 | New | 96 | 6 | 95 | 2304 |
| Qwen W3 | 6:2 | New | 96 | 12 | 95 | 2304 |
| Qwen W3 | 4:4 | Historical | 96 | 24 | 96 | 2304 |
| Qwen W1 | 8:0 | Historical | 96 | 0 | 85 | 2304 |
| Qwen W1 | 7:1 | New | 96 | 6 | 95 | 2304 |
| Qwen W1 | 6:2 | New | 96 | 12 | 96 | 2304 |
| Qwen W1 | 4:4 | Historical | 96 | 24 | 96 | 2304 |
| Qwen W4 | 8:0 | Historical | 96 | 0 | 83 | 2304 |
| Qwen W4 | 7:1 | New | 96 | 6 | 96 | 2304 |
| Qwen W4 | 6:2 | New | 96 | 12 | 96 | 2304 |
| Qwen W4 | 4:4 | Historical | 96 | 24 | 96 | 2304 |

Minority counts are application presentations per name. Every allocation uses eight distinct contexts and 48 application presentations per name. The deterministic masks are nested. M1 and M2 start independently from K. Unique counts identify names whose endpoint is uniquely determined by application labels at that allocation. These are design audits, not model performance.

Table 16: All six new fixed-dose child exports.

| World | Allocation | Updates | Examples | Seed | Weight digest |
|---|---|---|---|---|---|
| Qwen W3 | 7:1 | 2304 | 18432 | 61003 | `0b9abbac60` |
| Qwen W3 | 6:2 | 2304 | 18432 | 61003 | `567170c9cf` |
| Qwen W1 | 7:1 | 2304 | 18432 | 61001 | `292d30f812` |
| Qwen W1 | 6:2 | 2304 | 18432 | 61001 | `41526dd724` |
| Qwen W4 | 7:1 | 2304 | 18432 | 61004 | `d137c8a1a8` |
| Qwen W4 | 6:2 | 2304 | 18432 | 61004 | `9b23d3f618` |

Digest prefixes are identifiers only. Full checkpoint, parent, compiled-stream, configuration and training-trace hashes are retained in the source ledger and supplement. Later export reuse is not a further optimizer realization.

Table 17: Trained-name ordinary TEST levels. Every row uses 768 cases and 48 original history clusters.

| World | Path | Allocation | Accuracy | History 95% | Crossed 95% |
|---|---|---|---|---|---|
| Qwen W3 | B | K | 28.91 | [21.74, 36.72] | [19.57, 38.98] |
| Qwen W3 | B | 8:0 | 47.01 | [42.45, 51.69] | [41.42, 52.51] |
| Qwen W3 | B | 7:1 | 82.68 | [77.34, 87.50] | [75.47, 88.93] |
| Qwen W3 | B | 6:2 | 87.63 | [83.72, 91.15] | [81.55, 92.58] |
| Qwen W3 | B | 4:4 | 88.54 | [84.38, 92.32] | [82.36, 93.58] |
| Qwen W3 | S | K | 29.43 | [22.01, 37.50] | [19.97, 39.67] |
| Qwen W3 | S | 8:0 | 47.14 | [42.58, 51.69] | [41.52, 52.60] |
| Qwen W3 | S | 7:1 | 82.94 | [77.73, 87.76] | [75.76, 89.11] |
| Qwen W3 | S | 6:2 | 87.63 | [83.72, 91.15] | [81.55, 92.58] |
| Qwen W3 | S | 4:4 | 88.54 | [84.38, 92.32] | [82.36, 93.58] |
| Qwen W1 | B | K | 45.96 | [41.41, 50.39] | [36.85, 55.22] |
| Qwen W1 | B | 8:0 | 56.25 | [50.26, 62.24] | [49.27, 63.38] |
| Qwen W1 | B | 7:1 | 85.42 | [79.43, 90.62] | [77.96, 91.81] |
| Qwen W1 | B | 6:2 | 85.42 | [79.69, 90.49] | [77.91, 91.82] |
| Qwen W1 | B | 4:4 | 87.37 | [81.77, 92.19] | [80.30, 93.31] |
| Qwen W1 | S | K | 45.83 | [41.28, 50.26] | [36.76, 55.06] |
| Qwen W1 | S | 8:0 | 56.25 | [50.26, 62.24] | [49.27, 63.38] |
| Qwen W1 | S | 7:1 | 85.42 | [79.43, 90.62] | [77.96, 91.81] |
| Qwen W1 | S | 6:2 | 85.42 | [79.69, 90.49] | [77.91, 91.82] |
| Qwen W1 | S | 4:4 | 87.24 | [81.51, 92.19] | [80.06, 93.22] |
| Qwen W4 | B | K | 33.59 | [28.52, 38.80] | [24.47, 43.24] |
| Qwen W4 | B | 8:0 | 47.92 | [43.49, 52.08] | [42.63, 53.24] |
| Qwen W4 | B | 7:1 | 83.33 | [77.60, 88.54] | [76.12, 89.79] |
| Qwen W4 | B | 6:2 | 87.37 | [82.68, 91.54] | [80.65, 93.08] |
| Qwen W4 | B | 4:4 | 85.81 | [79.95, 91.02] | [78.36, 92.17] |
| Qwen W4 | S | K | 33.33 | [28.26, 38.54] | [24.22, 42.96] |
| Qwen W4 | S | 8:0 | 47.92 | [43.49, 52.08] | [42.63, 53.24] |
| Qwen W4 | S | 7:1 | 83.33 | [77.60, 88.54] | [76.06, 89.83] |
| Qwen W4 | S | 6:2 | 87.63 | [83.07, 91.80] | [81.06, 93.22] |
| Qwen W4 | S | 4:4 | 85.81 | [79.95, 91.02] | [78.36, 92.17] |

All intervals in these tables use the graded-extension reductions. The original endpoint models were not retrained; their previously reported estimates remain separate. Context counts are recorded in the ordinary-support tables.

Table 18: Qwen W3: all original trained TEST paired contrasts, in percentage points.

| World | Path | Contrast | Difference | History 95% | Crossed 95% |
|---|---|---|---|---|---|
| Qwen W3 | B | M1−M0 | +35.68 | [30.86, 40.49] | [29.57, 41.60] |
| Qwen W3 | B | M2−M1 | +4.95 | [1.95, 8.20] | [0.59, 9.51] |
| Qwen W3 | B | M4−M2 | +0.91 | [−1.43, 3.39] | [−2.91, 4.88] |
| Qwen W3 | B | M4−M0 | +41.54 | [36.98, 45.96] | [35.70, 47.23] |
| Qwen W3 | B | M0−K | +18.10 | [8.33, 27.21] | [5.91, 29.68] |
| Qwen W3 | B | M1−K | +53.78 | [42.84, 63.81] | [40.16, 66.36] |
| Qwen W3 | B | M2−K | +58.72 | [49.61, 67.19] | [46.68, 69.68] |
| Qwen W3 | B | M4−K | +59.64 | [50.00, 68.75] | [47.16, 71.31] |
| Qwen W3 | B | M1−M4 | -5.86 | [−8.46, −3.39] | [−10.22, −1.97] |
| Qwen W3 | B | M2−M4 | -0.91 | [−3.39, 1.43] | [−4.88, 2.91] |
| Qwen W3 | S | M1−M0 | +35.81 | [30.99, 40.62] | [29.65, 41.82] |
| Qwen W3 | S | M2−M1 | +4.69 | [1.82, 7.81] | [0.48, 9.03] |
| Qwen W3 | S | M4−M2 | +0.91 | [−1.43, 3.39] | [−2.91, 4.88] |
| Qwen W3 | S | M4−M0 | +41.41 | [36.85, 45.83] | [35.57, 47.11] |
| Qwen W3 | S | M0−K | +17.71 | [7.55, 27.08] | [5.34, 29.40] |
| Qwen W3 | S | M1−K | +53.52 | [42.45, 63.80] | [39.94, 66.27] |
| Qwen W3 | S | M2−K | +58.20 | [48.83, 66.93] | [46.04, 69.39] |
| Qwen W3 | S | M4−K | +59.11 | [49.22, 68.36] | [46.45, 70.99] |
| Qwen W3 | S | M1−M4 | -5.60 | [−8.20, −3.26] | [−9.93, −1.74] |
| Qwen W3 | S | M2−M4 | -0.91 | [−3.39, 1.43] | [−4.88, 2.91] |

The displayed 95% intervals are unadjusted. M1–M4 and M2–M4 are paired differences, not differences between interval endpoints. No nonsignificance-as-equivalence inference is made.

Table 19: Qwen W1: all original trained TEST paired contrasts, in percentage points.

| World | Path | Contrast | Difference | History 95% | Crossed 95% |
|---|---|---|---|---|---|
| Qwen W1 | B | M1−M0 | +29.17 | [25.26, 32.94] | [23.65, 34.43] |
| Qwen W1 | B | M2−M1 | +0.00 | [−1.95, 1.95] | [−3.25, 3.09] |
| Qwen W1 | B | M4−M2 | +1.95 | [0.00, 4.04] | [−1.03, 5.22] |
| Qwen W1 | B | M4−M0 | +31.12 | [26.95, 35.42] | [25.34, 36.68] |
| Qwen W1 | B | M0−K | +10.29 | [2.08, 18.10] | [−0.72, 20.97] |
| Qwen W1 | B | M1−K | +39.45 | [32.29, 46.35] | [28.36, 50.26] |
| Qwen W1 | B | M2−K | +39.45 | [32.29, 46.48] | [28.39, 50.23] |
| Qwen W1 | B | M4−K | +41.41 | [34.11, 48.18] | [30.54, 52.03] |
| Qwen W1 | B | M1−M4 | -1.95 | [−4.04, 0.13] | [−5.16, 1.00] |
| Qwen W1 | B | M2−M4 | -1.95 | [−4.04, 0.00] | [−5.22, 1.03] |
| Qwen W1 | S | M1−M0 | +29.17 | [25.26, 32.94] | [23.65, 34.43] |
| Qwen W1 | S | M2−M1 | +0.00 | [−1.95, 1.95] | [−3.25, 3.09] |
| Qwen W1 | S | M4−M2 | +1.82 | [−0.13, 3.91] | [−1.13, 5.02] |
| Qwen W1 | S | M4−M0 | +30.99 | [26.82, 35.29] | [25.25, 36.53] |
| Qwen W1 | S | M0−K | +10.42 | [2.21, 18.23] | [−0.60, 21.16] |
| Qwen W1 | S | M1−K | +39.58 | [32.29, 46.61] | [28.53, 50.44] |
| Qwen W1 | S | M2−K | +39.58 | [32.16, 46.61] | [28.36, 50.41] |
| Qwen W1 | S | M4−K | +41.41 | [33.98, 48.31] | [30.28, 52.05] |
| Qwen W1 | S | M1−M4 | -1.82 | [−3.91, 0.13] | [−4.97, 1.10] |
| Qwen W1 | S | M2−M4 | -1.82 | [−3.91, 0.13] | [−5.02, 1.13] |

The displayed 95% intervals are unadjusted. M1–M4 and M2–M4 are paired differences, not differences between interval endpoints. No nonsignificance-as-equivalence inference is made.

Table 20: Qwen W4: all original trained TEST paired contrasts, in percentage points.

| World | Path | Contrast | Difference | History 95% | Crossed 95% |
|---|---|---|---|---|---|
| Qwen W4 | B | M1−M0 | +35.42 | [31.77, 38.80] | [30.25, 40.20] |
| Qwen W4 | B | M2−M1 | +4.04 | [1.82, 6.51] | [0.40, 8.15] |
| Qwen W4 | B | M4−M2 | -1.56 | [−5.21, 1.82] | [−6.15, 2.96] |
| Qwen W4 | B | M4−M0 | +37.89 | [34.24, 41.28] | [32.47, 42.69] |
| Qwen W4 | B | M0−K | +14.32 | [8.20, 20.31] | [4.21, 23.87] |
| Qwen W4 | B | M1−K | +49.74 | [42.19, 57.29] | [38.45, 61.17] |
| Qwen W4 | B | M2−K | +53.78 | [47.01, 60.55] | [42.97, 64.45] |
| Qwen W4 | B | M4−K | +52.21 | [44.14, 59.77] | [40.49, 63.73] |
| Qwen W4 | B | M1−M4 | -2.47 | [−5.47, 0.65] | [−7.01, 2.14] |
| Qwen W4 | B | M2−M4 | +1.56 | [−1.82, 5.21] | [−2.96, 6.15] |
| Qwen W4 | S | M1−M0 | +35.42 | [31.77, 38.93] | [30.19, 40.22] |
| Qwen W4 | S | M2−M1 | +4.30 | [1.82, 7.03] | [0.40, 8.72] |
| Qwen W4 | S | M4−M2 | -1.82 | [−5.73, 1.69] | [−6.66, 2.81] |
| Qwen W4 | S | M4−M0 | +37.89 | [34.24, 41.28] | [32.47, 42.69] |
| Qwen W4 | S | M0−K | +14.58 | [8.59, 20.44] | [4.45, 24.07] |
| Qwen W4 | S | M1−K | +50.00 | [42.45, 57.43] | [38.75, 61.35] |
| Qwen W4 | S | M2−K | +54.30 | [47.53, 60.94] | [43.55, 64.83] |
| Qwen W4 | S | M4−K | +52.47 | [44.40, 60.03] | [40.78, 63.99] |
| Qwen W4 | S | M1−M4 | -2.47 | [−5.60, 0.65] | [−7.12, 2.19] |
| Qwen W4 | S | M2−M4 | +1.82 | [−1.69, 5.73] | [−2.81, 6.66] |

The displayed 95% intervals are unadjusted. M1–M4 and M2–M4 are paired differences, not differences between interval endpoints. No nonsignificance-as-equivalence inference is made.

Table 21: Nine-comparison adjusted adjacent contrasts on the original trained TEST panel.

| World | Path | Contrast | Difference | Adjusted H | Adjusted C |
|---|---|---|---|---|---|
| Qwen W3 | B | M1−M0 | +35.68 | [28.91, 42.58] | [26.93, 44.34] |
| Qwen W3 | B | M2−M1 | +4.95 | [0.65, 9.79] | [−1.05, 11.56] |
| Qwen W3 | B | M4−M2 | +0.91 | [−2.34, 4.43] | [−4.83, 6.84] |
| Qwen W3 | S | M1−M0 | +35.81 | [29.04, 42.74] | [27.05, 44.48] |
| Qwen W3 | S | M2−M1 | +4.69 | [0.52, 9.38] | [−1.22, 11.10] |
| Qwen W3 | S | M4−M2 | +0.91 | [−2.34, 4.43] | [−4.83, 6.84] |
| Qwen W1 | B | M1−M0 | +29.17 | [23.57, 34.64] | [20.91, 36.60] |
| Qwen W1 | B | M2−M1 | +0.00 | [−2.73, 2.73] | [−4.98, 4.61] |
| Qwen W1 | B | M4−M2 | +1.95 | [−0.81, 5.08] | [−2.48, 6.93] |
| Qwen W1 | S | M1−M0 | +29.17 | [23.57, 34.64] | [20.91, 36.60] |
| Qwen W1 | S | M2−M1 | +0.00 | [−2.73, 2.73] | [−4.98, 4.61] |
| Qwen W1 | S | M4−M2 | +1.82 | [−0.91, 4.72] | [−2.61, 6.88] |
| Qwen W4 | B | M1−M0 | +35.42 | [30.21, 40.10] | [27.80, 41.94] |
| Qwen W4 | B | M2−M1 | +4.04 | [0.78, 7.55] | [−1.40, 10.20] |
| Qwen W4 | B | M4−M2 | -1.56 | [−6.93, 3.26] | [−8.85, 5.10] |
| Qwen W4 | S | M1−M0 | +35.42 | [30.08, 39.97] | [27.68, 41.95] |
| Qwen W4 | S | M2−M1 | +4.30 | [0.65, 8.23] | [−1.46, 10.81] |
| Qwen W4 | S | M4−M2 | -1.82 | [−7.55, 3.12] | [−9.38, 4.96] |

Each interval has confidence $1 - 0.05/9$. The family is the three adjacent comparisons in each of three worlds, separately retained for each scoring path. This adjustment does not cover secondary populations, restrictions, or all displayed comparisons.

Table 22: Equal-world Qwen effects and the separate three-world sensitivity.

| Path | Contrast | Mean | World t 95% |
|---|---|---|---|
| B | M1−M0 | +33.42 | $[24.26, 42.58]$ |
| B | M2−M1 | +2.99 | $[-3.55, 9.54]$ |
| B | M4−M2 | +0.43 | $[-4.05, 4.92]$ |
| S | M1−M0 | +33.46 | $[24.21, 42.72]$ |
| S | M2−M1 | +2.99 | $[-3.47, 9.46]$ |
| S | M4−M2 | +0.30 | $[-4.41, 5.02]$ |

The three world effects are averaged equally. The interval uses two degrees of freedom. It is separate from within-world history/crossed uncertainty and does not establish across-optimizer-seed stability.

Table 23: Qwen W3: ordinary DIRECT levels for every original population and relevant rendering.

| Support | Path | $n/h/c$ | K | 8:0 | 7:1 | 6:2 | 4:4 |
|---|---|---|---|---|---|---|---|
| T test | B | 768/48/141 | 28.91 | 47.01 | 82.68 | 87.63 | 88.54 |
| F test | B | 1536/96/157 | 29.88 | 41.41 | 48.31 | 46.88 | 54.30 |
| F test alt. | B | 1536/96/157 | 30.53 | 39.78 | 48.57 | 46.94 | 53.65 |
| F+ test | B | 384/24/107 | 30.21 | 35.16 | 46.09 | 48.44 | 48.70 |
| F+ test alt. | B | 384/24/107 | 30.47 | 34.11 | 46.61 | 48.96 | 48.70 |
| F val | B | 384/24/107 | 22.66 | 39.84 | 46.09 | 34.90 | 45.31 |
| T test alt. | B | 768/48/141 | 29.30 | 47.01 | 82.68 | 87.37 | 88.41 |
| T val | B | 768/48/146 | 27.60 | 46.22 | 83.59 | 86.33 | 87.37 |
| T test | S | 768/48/141 | 29.43 | 47.14 | 82.94 | 87.63 | 88.54 |
| F test | S | 1536/96/157 | 30.08 | 41.54 | 48.70 | 46.68 | 54.10 |
| F test alt. | S | 1536/96/157 | 30.27 | 39.45 | 48.70 | 47.01 | 53.26 |
| F+ test | S | 384/24/107 | 29.95 | 34.64 | 46.35 | 48.70 | 48.70 |
| F+ test alt. | S | 384/24/107 | 30.47 | 34.11 | 47.14 | 48.96 | 48.70 |
| F val | S | 384/24/107 | 22.66 | 40.10 | 45.83 | 35.16 | 45.05 |
| T test alt. | S | 768/48/141 | 29.30 | 46.88 | 82.81 | 87.63 | 88.41 |
| T val | S | 768/48/146 | 27.34 | 46.09 | 83.46 | 86.46 | 87.11 |

Values are percent correct. T denotes application-trained names, F the original fact-only population, and F+ the additional fact-only test population. Alternate rendering is marked alt. Validation and test stay separate. Recall has no visible context; $c = 0$ and its sensitivity resamples histories only. Full per-cell intervals, paired effects, validity counts and compatibility outcomes are in the complete source exports.

Table 24: Qwen W3: recall, supplied-true and supplied-conflict levels in every original component cell.

| Support | Path | $n/h/c$ | K | 8:0 | 7:1 | 6:2 | 4:4 |
|---|---|---|---|---|---|---|---|
| True F test | B | 1536/96/157 | 92.32 | 97.14 | 96.94 | 97.85 | 96.03 |
| True F+ test | B | 384/24/107 | 91.93 | 98.44 | 98.96 | 96.88 | 97.40 |
| True F val | B | 384/24/107 | 91.15 | 95.83 | 94.79 | 97.40 | 96.61 |
| True T test | B | 768/48/141 | 94.01 | 97.79 | 98.44 | 99.09 | 97.66 |
| True T val | B | 768/48/146 | 88.54 | 95.18 | 95.83 | 95.18 | 95.31 |
| Conflict T test | B | 768/48/141 | 93.75 | 96.74 | 97.01 | 97.53 | 95.31 |
| Conflict T val | B | 768/48/146 | 89.45 | 95.70 | 95.83 | 97.53 | 96.48 |
| Recall All facts | B | 384/192/0 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Recall All facts alt. | B | 384/192/0 | 99.74 | 99.74 | 99.74 | 100.00 | 99.74 |
| True F test | S | 1536/96/157 | 92.25 | 97.20 | 96.94 | 97.85 | 96.09 |
| True F+ test | S | 384/24/107 | 91.93 | 98.44 | 98.96 | 96.88 | 97.40 |
| True F val | S | 384/24/107 | 91.15 | 95.83 | 94.79 | 97.40 | 96.61 |
| True T test | S | 768/48/141 | 94.01 | 97.79 | 98.44 | 99.09 | 97.66 |
| True T val | S | 768/48/146 | 88.41 | 95.18 | 95.83 | 95.18 | 95.31 |
| Conflict T test | S | 768/48/141 | 93.75 | 96.74 | 96.88 | 97.40 | 95.31 |
| Conflict T val | S | 768/48/146 | 89.45 | 95.70 | 95.83 | 97.53 | 96.35 |
| Recall All facts | S | 384/192/0 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Recall All facts alt. | S | 384/192/0 | 99.74 | 99.74 | 99.74 | 100.00 | 99.74 |

Values are percent correct. T denotes application-trained names, F the original fact-only population, and F+ the additional fact-only test population. Alternate rendering is marked alt. Validation and test stay separate. Recall has no visible context; $c = 0$ and its sensitivity resamples histories only. Full per-cell intervals, paired effects, validity counts and compatibility outcomes are in the complete source exports.

Table 25: Qwen W1: ordinary DIRECT levels for every original population and relevant rendering.

| Support | Path | $n/h/c$ | K | 8:0 | 7:1 | 6:2 | 4:4 |
|---|---|---|---|---|---|---|---|
| T test | B | 768/48/146 | 45.96 | 56.25 | 85.42 | 85.42 | 87.37 |
| F test | B | 1536/96/171 | 45.44 | 46.61 | 57.36 | 58.98 | 64.19 |
| F test alt. | B | 1536/96/171 | 45.51 | 46.22 | 57.49 | 60.29 | 65.04 |
| F+ test | B | 384/24/112 | 38.02 | 48.96 | 69.01 | 64.32 | 67.97 |
| F+ test alt. | B | 384/24/112 | 38.28 | 49.48 | 67.97 | 63.80 | 67.71 |
| F val | B | 384/24/110 | 42.19 | 32.29 | 46.88 | 58.07 | 57.03 |
| T test alt. | B | 768/48/146 | 45.96 | 56.77 | 85.42 | 85.55 | 87.24 |
| T val | B | 768/48/155 | 44.14 | 59.77 | 86.85 | 88.02 | 89.84 |
| T test | S | 768/48/146 | 45.83 | 56.25 | 85.42 | 85.42 | 87.24 |
| F test | S | 1536/96/171 | 45.18 | 46.74 | 57.29 | 59.11 | 64.19 |
| F test alt. | S | 1536/96/171 | 46.03 | 46.09 | 57.75 | 60.16 | 64.84 |
| F+ test | S | 384/24/112 | 38.02 | 48.96 | 68.75 | 64.32 | 68.23 |
| F+ test alt. | S | 384/24/112 | 38.28 | 49.48 | 68.23 | 63.80 | 67.45 |
| F val | S | 384/24/110 | 42.19 | 32.29 | 46.88 | 58.07 | 57.03 |
| T test alt. | S | 768/48/146 | 45.44 | 56.51 | 85.42 | 85.42 | 87.24 |
| T val | S | 768/48/155 | 44.27 | 59.77 | 86.85 | 87.89 | 89.84 |

Values are percent correct. T denotes application-trained names, F the original fact-only population, and F+ the additional fact-only test population. Alternate rendering is marked alt. Validation and test stay separate. Recall has no visible context; $c = 0$ and its sensitivity resamples histories only. Full per-cell intervals, paired effects, validity counts and compatibility outcomes are in the complete source exports.

Table 26: Qwen W1: recall, supplied-true and supplied-conflict levels in every original component cell.

| Support | Path | $n/h/c$ | K | 8:0 | 7:1 | 6:2 | 4:4 |
|---|---|---|---|---|---|---|---|
| True F test | B | 1536/96/171 | 94.08 | 95.12 | 94.99 | 93.16 | 95.12 |
| True F+ test | B | 384/24/112 | 95.83 | 97.40 | 96.35 | 97.66 | 96.88 |
| True F val | B | 384/24/110 | 91.67 | 97.40 | 96.61 | 93.23 | 95.57 |
| True T test | B | 768/48/146 | 94.14 | 97.01 | 95.83 | 93.75 | 96.48 |
| True T val | B | 768/48/155 | 92.06 | 98.05 | 97.79 | 96.09 | 97.01 |
| Conflict T test | B | 768/48/146 | 91.67 | 96.35 | 96.35 | 95.70 | 96.35 |
| Conflict T val | B | 768/48/155 | 85.03 | 95.57 | 95.05 | 94.01 | 93.49 |
| Recall All facts | B | 384/192/0 | 100.00 | 100.00 | 100.00 | 100.00 | 98.70 |
| Recall All facts alt. | B | 384/192/0 | 100.00 | 100.00 | 100.00 | 100.00 | 98.70 |
| True F test | S | 1536/96/171 | 93.82 | 95.18 | 94.99 | 93.10 | 95.12 |
| True F+ test | S | 384/24/112 | 95.83 | 97.40 | 96.35 | 97.92 | 96.88 |
| True F val | S | 384/24/110 | 91.67 | 97.40 | 96.61 | 93.23 | 95.57 |
| True T test | S | 768/48/146 | 94.14 | 97.14 | 95.70 | 93.75 | 96.48 |
| True T val | S | 768/48/155 | 92.32 | 98.18 | 97.79 | 96.35 | 97.01 |
| Conflict T test | S | 768/48/146 | 91.67 | 96.35 | 96.35 | 95.70 | 96.35 |
| Conflict T val | S | 768/48/155 | 85.29 | 95.31 | 95.05 | 94.01 | 93.62 |
| Recall All facts | S | 384/192/0 | 100.00 | 100.00 | 100.00 | 100.00 | 98.70 |
| Recall All facts alt. | S | 384/192/0 | 100.00 | 100.00 | 100.00 | 100.00 | 98.70 |

Values are percent correct. T denotes application-trained names, F the original fact-only population, and F+ the additional fact-only test population. Alternate rendering is marked alt. Validation and test stay separate. Recall has no visible context; $c = 0$ and its sensitivity resamples histories only. Full per-cell intervals, paired effects, validity counts and compatibility outcomes are in the complete source exports.

Table 27: Qwen W4: ordinary DIRECT levels for every original population and relevant rendering.

| Support | Path | $n/h/c$ | K | 8:0 | 7:1 | 6:2 | 4:4 |
|---|---|---|---|---|---|---|---|
| T test | B | 768/48/141 | 33.59 | 47.92 | 83.33 | 87.37 | 85.81 |
| F test | B | 1536/96/155 | 37.63 | 55.53 | 73.89 | 72.79 | 69.60 |
| F test alt. | B | 1536/96/155 | 36.72 | 55.34 | 75.72 | 72.79 | 71.35 |
| F+ test | B | 384/24/111 | 40.62 | 57.55 | 69.01 | 69.27 | 66.67 |
| F+ test alt. | B | 384/24/111 | 41.41 | 59.38 | 70.31 | 67.97 | 66.41 |
| F val | B | 384/24/122 | 27.08 | 42.97 | 57.81 | 64.58 | 62.76 |
| T test alt. | B | 768/48/141 | 32.94 | 48.44 | 82.94 | 86.72 | 85.29 |
| T val | B | 768/48/142 | 29.04 | 46.74 | 81.64 | 85.03 | 85.81 |
| T test | S | 768/48/141 | 33.33 | 47.92 | 83.33 | 87.63 | 85.81 |
| F test | S | 1536/96/155 | 37.24 | 55.60 | 74.09 | 72.79 | 69.40 |
| F test alt. | S | 1536/96/155 | 36.91 | 56.05 | 75.59 | 72.66 | 71.48 |
| F+ test | S | 384/24/111 | 40.10 | 57.29 | 68.49 | 69.53 | 66.67 |
| F+ test alt. | S | 384/24/111 | 41.41 | 59.64 | 70.57 | 67.97 | 66.15 |
| F val | S | 384/24/122 | 27.08 | 42.97 | 57.81 | 64.06 | 62.24 |
| T test alt. | S | 768/48/141 | 32.55 | 48.44 | 82.94 | 86.85 | 85.42 |
| T val | S | 768/48/142 | 29.04 | 46.74 | 81.51 | 85.03 | 85.68 |

Values are percent correct. T denotes application-trained names, F the original fact-only population, and F+ the additional fact-only test population. Alternate rendering is marked alt. Validation and test stay separate. Recall has no visible context; $c = 0$ and its sensitivity resamples histories only. Full per-cell intervals, paired effects, validity counts and compatibility outcomes are in the complete source exports.

Table 28: Qwen W4: recall, supplied-true and supplied-conflict levels in every original component cell.

| Support | Path | $n/h/c$ | K | 8:0 | 7:1 | 6:2 | 4:4 |
|---|---|---|---|---|---|---|---|
| True F test | B | 1536/96/155 | 93.10 | 96.22 | 94.53 | 95.31 | 96.48 |
| True F+ test | B | 384/24/111 | 91.93 | 95.31 | 94.01 | 94.27 | 95.57 |
| True F val | B | 384/24/122 | 91.67 | 95.83 | 93.49 | 94.79 | 97.40 |
| True T test | B | 768/48/141 | 92.32 | 97.01 | 96.61 | 94.27 | 97.01 |
| True T val | B | 768/48/142 | 92.19 | 94.53 | 94.14 | 94.14 | 95.83 |
| Conflict T test | B | 768/48/141 | 92.32 | 95.31 | 80.08 | 86.46 | 90.36 |
| Conflict T val | B | 768/48/142 | 89.06 | 92.06 | 80.08 | 84.51 | 89.58 |
| Recall All facts | B | 384/192/0 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Recall All facts alt. | B | 384/192/0 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| True F test | S | 1536/96/155 | 93.10 | 96.22 | 94.60 | 95.31 | 96.55 |
| True F+ test | S | 384/24/111 | 91.93 | 95.31 | 94.01 | 94.27 | 95.57 |
| True F val | S | 384/24/122 | 91.67 | 95.83 | 93.49 | 94.79 | 97.40 |
| True T test | S | 768/48/141 | 92.32 | 97.01 | 96.61 | 94.53 | 97.01 |
| True T val | S | 768/48/142 | 92.19 | 94.66 | 94.27 | 94.14 | 95.83 |
| Conflict T test | S | 768/48/141 | 92.32 | 95.31 | 80.21 | 86.46 | 90.49 |
| Conflict T val | S | 768/48/142 | 89.06 | 92.19 | 80.08 | 84.51 | 89.45 |
| Recall All facts | S | 384/192/0 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Recall All facts alt. | S | 384/192/0 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

Values are percent correct. T denotes application-trained names, F the original fact-only population, and F+ the additional fact-only test population. Alternate rendering is marked alt. Validation and test stay separate. Recall has no visible context; $c = 0$ and its sensitivity resamples histories only. Full per-cell intervals, paired effects, validity counts and compatibility outcomes are in the complete source exports.

Table 29: Final training-prompt fit for all six new children.

| World | Path | Allocation | Majority $k/n$ | % | Minority $k/n$ | % |
|---|---|---|---|---|---|---|
| Qwen W3 | B | 7:1 | 671/672 | 99.85 | 92/96 | 95.83 |
| Qwen W3 | B | 6:2 | 575/576 | 99.83 | 191/192 | 99.48 |
| Qwen W3 | S | 7:1 | 671/672 | 99.85 | 92/96 | 95.83 |
| Qwen W3 | S | 6:2 | 575/576 | 99.83 | 191/192 | 99.48 |
| Qwen W1 | B | 7:1 | 668/672 | 99.40 | 92/96 | 95.83 |
| Qwen W1 | B | 6:2 | 573/576 | 99.48 | 189/192 | 98.44 |
| Qwen W1 | S | 7:1 | 668/672 | 99.40 | 92/96 | 95.83 |
| Qwen W1 | S | 6:2 | 573/576 | 99.48 | 189/192 | 98.44 |
| Qwen W4 | B | 7:1 | 669/672 | 99.55 | 90/96 | 93.75 |
| Qwen W4 | B | 6:2 | 572/576 | 99.31 | 191/192 | 99.48 |
| Qwen W4 | S | 7:1 | 669/672 | 99.55 | 90/96 | 93.75 |
| Qwen W4 | S | 6:2 | 572/576 | 99.31 | 191/192 | 99.48 |

These are distinct assigned application prompts in one fixed training rendering. They are descriptive diagnostics, not checkpoint selection or additional held-out evidence. Each original context was presented six times during the unchanged training schedule.

Table 30: Original fact-only TEST gains and the separate additional fact-only population.

| World | Path | Pop. | Contrast | Difference | History 95% | Crossed 95% |
|---|---|---|---|---|---|---|
| Qwen W3 | B | F | M1−M0 | +6.90 | [3.45, 10.29] | [1.74, 12.15] |
| Qwen W3 | B | F | M1−K | +18.42 | [8.07, 28.58] | [6.68, 29.82] |
| Qwen W3 | B | F+ | M1−K | +15.89 | [−6.25, 37.76] | [−7.65, 40.11] |
| Qwen W3 | S | F | M1−M0 | +7.16 | [3.71, 10.55] | [1.95, 12.44] |
| Qwen W3 | S | F | M1−K | +18.62 | [8.27, 28.71] | [6.83, 30.01] |
| Qwen W3 | S | F+ | M1−K | +16.41 | [−5.73, 38.54] | [−7.40, 40.57] |
| Qwen W1 | B | F | M1−M0 | +10.74 | [6.38, 15.10] | [5.31, 16.59] |
| Qwen W1 | B | F | M1−K | +11.91 | [4.43, 19.34] | [2.11, 21.82] |
| Qwen W1 | B | F+ | M1−K | +30.99 | [17.71, 42.97] | [13.41, 48.63] |
| Qwen W1 | S | F | M1−M0 | +10.55 | [6.25, 14.84] | [5.15, 16.31] |
| Qwen W1 | S | F | M1−K | +12.11 | [4.62, 19.47] | [2.32, 22.07] |
| Qwen W1 | S | F+ | M1−K | +30.73 | [17.19, 42.98] | [12.99, 48.46] |
| Qwen W4 | B | F | M1−M0 | +18.36 | [13.80, 22.98] | [12.67, 24.21] |
| Qwen W4 | B | F | M1−K | +36.26 | [29.30, 42.97] | [27.44, 45.23] |
| Qwen W4 | B | F+ | M1−K | +28.39 | [12.24, 43.75] | [9.13, 47.01] |
| Qwen W4 | S | F | M1−M0 | +18.49 | [14.00, 22.98] | [12.85, 24.29] |
| Qwen W4 | S | F | M1−K | +36.85 | [29.95, 43.55] | [28.04, 45.70] |
| Qwen W4 | S | F+ | M1−K | +28.39 | [12.50, 43.23] | [9.43, 46.59] |

F uses the original 1,536-case fact-only TEST panel. F+ uses the additional 384-case fact-only TEST panel. These are paired secondary comparisons, unadjusted for the nine-comparison primary family. In particular, a weak F+ result is not replaced by the stronger F result.

Table 31: Supplied-conflict accuracy costs and changes on the 768 original trained-name TEST cases.

| World | Path | Contrast | Difference | History 95% | Crossed 95% |
|---|---|---|---|---|---|
| Qwen W3 | B | M1−M0 | +0.26 | [−1.17, 1.69] | [−2.28, 2.87] |
| Qwen W3 | B | M1−K | +3.26 | [1.56, 4.95] | [−0.21, 7.08] |
| Qwen W3 | B | M2−M1 | +0.52 | [−0.78, 1.82] | [−1.77, 3.16] |
| Qwen W3 | S | M1−M0 | +0.13 | [−1.30, 1.56] | [−2.66, 2.83] |
| Qwen W3 | S | M1−K | +3.12 | [1.43, 4.82] | [−0.35, 6.92] |
| Qwen W3 | S | M2−M1 | +0.52 | [−0.78, 1.82] | [−1.71, 3.09] |
| Qwen W1 | B | M1−M0 | +0.00 | [−1.43, 1.30] | [−3.15, 2.96] |
| Qwen W1 | B | M1−K | +4.69 | [2.08, 7.42] | [−0.25, 9.88] |
| Qwen W1 | B | M2−M1 | -0.65 | [−2.21, 0.91] | [−3.62, 2.59] |
| Qwen W1 | S | M1−M0 | +0.00 | [−1.43, 1.30] | [−3.15, 2.96] |
| Qwen W1 | S | M1−K | +4.69 | [1.95, 7.55] | [−0.26, 9.97] |
| Qwen W1 | S | M2−M1 | -0.65 | [−2.21, 0.91] | [−3.62, 2.59] |
| Qwen W4 | B | M1−M0 | -15.23 | [−21.35, −9.77] | [−23.01, −8.43] |
| Qwen W4 | B | M1−K | -12.24 | [−19.27, −5.86] | [−21.33, −3.81] |
| Qwen W4 | B | M2−M1 | +6.38 | [2.34, 10.55] | [0.53, 12.97] |
| Qwen W4 | S | M1−M0 | -15.10 | [−21.22, −9.77] | [−22.89, −8.30] |
| Qwen W4 | S | M1−K | -12.11 | [−19.01, −5.86] | [−21.26, −3.58] |
| Qwen W4 | S | M2−M1 | +6.25 | [2.47, 10.29] | [0.49, 12.75] |

Correctness uses the fixed supplied-endpoint target. Intervals are unadjusted secondary comparisons. The adverse W4 M1–M0 result persists on both numerical paths. It does not establish a universal cost or which internal memory source produced an error. The separate optional Phi error-category decomposition was disabled.

Table 32: ANSWER_UNSEEN absolute levels, with K retained beside every application-trained condition.

| Block | Path | $n/h/c$ | K | 8:0 | 7:1 | 6:2 | 4:4 |
|---|---|---|---|---|---|---|---|
| Qwen W3 | B | 152/25/57 | 41.47 | 7.27 | 36.90 | 46.50 | 46.30 |
| Qwen W3 | S | 152/25/57 | 41.47 | 7.27 | 37.40 | 46.50 | 46.30 |
| Qwen W1 | B | 106/17/43 | 50.88 | 5.59 | 20.05 | 12.70 | 23.92 |
| Qwen W1 | S | 106/17/43 | 51.47 | 5.59 | 20.05 | 12.70 | 23.33 |
| Qwen W4 | B | 110/19/48 | 35.79 | 3.51 | 21.58 | 30.09 | 25.70 |
| Qwen W4 | S | 110/19/48 | 34.91 | 3.51 | 21.58 | 31.27 | 25.70 |
| Llama W3 | B | 152/25/57 | 53.17 | 18.50 | – | – | 48.70 |
| Llama W3 | S | 152/25/57 | 53.17 | 18.50 | – | – | 48.70 |
| Phi R1 | B | 152/25/57 | 36.60 | 4.50 | – | – | 11.93 |
| Phi R1 | S | 152/25/57 | 36.60 | 4.50 | – | – | 11.93 |
| Phi R2 | B | 156/26/57 | 63.21 | 0.00 | – | – | 25.26 |
| Phi R2 | S | 156/26/57 | 63.21 | 0.00 | – | – | 25.26 |

This post hoc P/S-only subset excludes each name's union of M0/M4 application labels. The labels are familiar globally. Accuracy averages histories equally, so raw correct count divided by cases need not equal the displayed value. Llama/Phi B and S rows repeat the same cached observations. Dashes denote levels never trained for those blocks. ANSWER_UNSEEN is not FACT_ONLY_UNEXPOSED.

Table 33: ANSWER_UNSEEN sibling and parent contrasts for every original block.

| Block | Path | Contrast | Difference | History 95% | Crossed 95% |
|---|---|---|---|---|---|
| Qwen W3 | B | M4−M0 | +39.03 | $[23.73, 53.93]$ | $[20.88, 59.58]$ |
| Qwen W3 | B | M4−K | +4.83 | $[-19.33, 29.67]$ | $[-21.08, 37.30]$ |
| Qwen W3 | S | M4−M0 | +39.03 | $[23.73, 53.93]$ | $[20.88, 59.58]$ |
| Qwen W3 | S | M4−K | +4.83 | $[-19.33, 29.67]$ | $[-21.08, 37.30]$ |
| Qwen W1 | B | M4−M0 | +18.33 | $[5.78, 32.45]$ | $[1.85, 39.74]$ |
| Qwen W1 | B | M4−K | -26.96 | $[-40.15, -15.64]$ | $[-51.67, 2.50]$ |
| Qwen W1 | S | M4−M0 | +17.75 | $[5.44, 31.86]$ | $[1.39, 39.09]$ |
| Qwen W1 | S | M4−K | -28.14 | $[-41.27, -16.57]$ | $[-52.78, 1.31]$ |
| Qwen W4 | B | M4−M0 | +22.19 | $[9.47, 36.89]$ | $[6.67, 39.87]$ |
| Qwen W4 | B | M4−K | -10.09 | $[-27.06, 8.60]$ | $[-36.52, 17.67]$ |
| Qwen W4 | S | M4−M0 | +22.19 | $[9.47, 36.89]$ | $[6.67, 39.87]$ |
| Qwen W4 | S | M4−K | -9.21 | $[-26.23, 9.47]$ | $[-35.64, 18.41]$ |
| Llama W3 | B | M4−M0 | +30.20 | $[12.50, 47.50]$ | $[9.95, 52.38]$ |
| Llama W3 | B | M4−K | -4.47 | $[-29.50, 20.27]$ | $[-35.87, 24.64]$ |
| Llama W3 | S | M4−M0 | +30.20 | $[12.50, 47.50]$ | $[9.95, 52.38]$ |
| Llama W3 | S | M4−K | -4.47 | $[-29.50, 20.27]$ | $[-35.87, 24.64]$ |
| Phi R1 | B | M4−M0 | +7.43 | $[-2.77, 17.50]$ | $[-5.32, 19.94]$ |
| Phi R1 | B | M4−K | -24.67 | $[-42.40, -7.73]$ | $[-46.16, -2.36]$ |
| Phi R1 | S | M4−M0 | +7.43 | $[-2.77, 17.50]$ | $[-5.32, 19.94]$ |
| Phi R1 | S | M4−K | -24.67 | $[-42.40, -7.73]$ | $[-46.16, -2.36]$ |
| Phi R2 | B | M4−M0 | +25.26 | $[13.46, 37.95]$ | $[11.11, 43.23]$ |
| Phi R2 | B | M4−K | -37.95 | $[-57.98, -17.60]$ | $[-63.20, -7.94]$ |
| Phi R2 | S | M4−M0 | +25.26 | $[13.46, 37.95]$ | $[11.11, 43.23]$ |
| Phi R2 | S | M4−K | -37.95 | $[-57.98, -17.60]$ | $[-63.20, -7.94]$ |

Differences are percentage points on exactly the same within-block subset. Both uncertainty schemes are retained, including unresolved and negative findings. These secondary comparisons are unadjusted.

Table 34: Graded ANSWER_UNSEEN contrasts in all three Qwen worlds.

| World | Path | Contrast | Difference | History 95% | Crossed 95% |
|---|---|---|---|---|---|
| Qwen W3 | B | M1−M0 | +29.63 | $[16.33, 42.87]$ | $[11.81, 47.79]$ |
| Qwen W3 | B | M1−K | -4.57 | $[-28.13, 19.33]$ | $[-32.07, 27.33]$ |
| Qwen W3 | B | M2−K | +5.03 | $[-14.30, 25.57]$ | $[-16.15, 32.82]$ |
| Qwen W3 | S | M1−M0 | +30.13 | $[16.73, 43.47]$ | $[12.49, 48.43]$ |
| Qwen W3 | S | M1−K | -4.07 | $[-27.50, 19.63]$ | $[-31.48, 27.70]$ |
| Qwen W3 | S | M2−K | +5.03 | $[-14.30, 25.57]$ | $[-16.15, 32.82]$ |
| Qwen W1 | B | M1−M0 | +14.46 | $[3.68, 25.74]$ | $[0.00, 33.50]$ |
| Qwen W1 | B | M1−K | -30.83 | $[-45.29, -16.62]$ | $[-57.92, 1.76]$ |
| Qwen W1 | B | M2−K | -38.19 | $[-50.44, -25.74]$ | $[-61.34, -12.71]$ |
| Qwen W1 | S | M1−M0 | +14.46 | $[3.68, 25.74]$ | $[0.00, 33.50]$ |
| Qwen W1 | S | M1−K | -31.42 | $[-45.83, -17.16]$ | $[-58.43, 0.76]$ |
| Qwen W1 | S | M2−K | -38.77 | $[-50.98, -26.47]$ | $[-61.75, -13.34]$ |
| Qwen W4 | B | M1−M0 | +18.07 | $[5.35, 32.89]$ | $[2.78, 35.24]$ |
| Qwen W4 | B | M1−K | -14.21 | $[-26.93, -0.88]$ | $[-36.16, 4.59]$ |
| Qwen W4 | B | M2−K | -5.70 | $[-22.24, 9.74]$ | $[-28.25, 19.79]$ |
| Qwen W4 | S | M1−M0 | +18.07 | $[5.35, 32.89]$ | $[2.78, 35.24]$ |
| Qwen W4 | S | M1−K | -13.33 | $[-26.14, 0.18]$ | $[-35.19, 5.33]$ |
| Qwen W4 | S | M2−K | -3.64 | $[-20.48, 11.93]$ | $[-25.93, 21.78]$ |

The W1 M1–M0 crossed lower bound is exactly zero. It is not a resolved positive effect under both schemes. All remaining graded subset contrasts, including differences from M4, remain in the full source exports.

Table 35: All-level common-identifiable accuracy and its whole-pair sensitivity.

| World / support | Path | $n/h/c$ | K | 8:0 | 7:1 | 6:2 | 4:4 |
|---|---|---|---|---|---|---|---|
| Qwen W3 Names | B | 584/48/141 | 27.99 | 39.19 | 81.51 | 87.76 | 87.11 |
| Qwen W3 Names | S | 584/48/141 | 28.52 | 39.19 | 81.64 | 87.76 | 87.11 |
| Qwen W3 Pairs | B | 400/25/102 | 26.25 | 50.25 | 87.50 | 91.50 | 94.25 |
| Qwen W3 Pairs | S | 400/25/102 | 27.25 | 50.25 | 87.75 | 91.50 | 94.25 |
| Qwen W1 Names | B | 680/48/146 | 45.83 | 54.69 | 84.90 | 86.07 | 86.85 |
| Qwen W1 Names | S | 680/48/146 | 45.57 | 54.69 | 84.90 | 86.07 | 86.72 |
| Qwen W1 Pairs | B | 592/37/131 | 42.23 | 62.16 | 90.20 | 91.05 | 93.07 |
| Qwen W1 Pairs | S | 592/37/131 | 41.89 | 62.16 | 90.20 | 91.05 | 92.91 |
| Qwen W4 Names | B | 664/48/141 | 32.94 | 45.44 | 82.81 | 86.98 | 86.33 |
| Qwen W4 Names | S | 664/48/141 | 32.81 | 45.44 | 82.81 | 87.24 | 86.33 |
| Qwen W4 Pairs | B | 560/35/132 | 33.39 | 52.68 | 89.64 | 93.57 | 91.25 |
| Qwen W4 Pairs | S | 560/35/132 | 33.21 | 52.68 | 89.64 | 93.93 | 91.25 |

Names retains entities whose endpoint is uniquely identified by application labels at every evaluated allocation. Pairs requires both original pair members to qualify. This is the new all-level restriction, not a replacement for the original-comparison identifiability strata.

Table 36: Common-identifiable names: graded benefit, parent comparison and difference from M4.

| World | Path | Contrast | Difference | History 95% | Crossed 95% |
|---|---|---|---|---|---|
| Qwen W3 | B | M1−M0 | +42.32 | [36.20, 48.31] | [34.02, 50.87] |
| Qwen W3 | B | M1−K | +53.52 | [42.06, 64.32] | [39.16, 66.49] |
| Qwen W3 | B | M1−M4 | -5.60 | [−8.59, −2.73] | [−10.65, −1.20] |
| Qwen W3 | S | M1−M0 | +42.45 | [36.33, 48.44] | [34.11, 51.05] |
| Qwen W3 | S | M1−K | +53.12 | [41.67, 64.06] | [38.70, 66.33] |
| Qwen W3 | S | M1−M4 | -5.47 | [−8.46, −2.60] | [−10.60, −1.10] |
| Qwen W1 | B | M1−M0 | +30.21 | [26.04, 34.11] | [23.77, 36.55] |
| Qwen W1 | B | M1−K | +39.06 | [31.77, 46.09] | [27.96, 49.86] |
| Qwen W1 | B | M1−M4 | -1.95 | [−4.04, −0.13] | [−5.17, 1.16] |
| Qwen W1 | S | M1−M0 | +30.21 | [26.04, 34.11] | [23.77, 36.55] |
| Qwen W1 | S | M1−K | +39.32 | [32.03, 46.35] | [28.26, 50.15] |
| Qwen W1 | S | M1−M4 | -1.82 | [−3.78, 0.00] | [−5.01, 1.24] |
| Qwen W4 | B | M1−M0 | +37.37 | [33.33, 41.15] | [31.20, 43.30] |
| Qwen W4 | B | M1−K | +49.87 | [42.19, 57.42] | [38.65, 61.13] |
| Qwen W4 | B | M1−M4 | -3.52 | [−7.16, 0.00] | [−8.66, 1.28] |
| Qwen W4 | S | M1−M0 | +37.37 | [33.33, 41.28] | [31.17, 43.33] |
| Qwen W4 | S | M1−K | +50.00 | [42.45, 57.42] | [38.77, 61.24] |
| Qwen W4 | S | M1−M4 | -3.52 | [−7.29, 0.00] | [−8.73, 1.31] |

All differences are percentage points with unadjusted conditional intervals. Restriction weakens endpoint ambiguity as a sole explanation. It does not equalize optimization difficulty or establish a component mechanism.

Table 37: Whole-pair common-identifiable support: graded benefit, parent comparison and difference from M4.

| World | Path | Contrast | Difference | History 95% | Crossed 95% |
|---|---|---|---|---|---|
| Qwen W3 | B | M1−M0 | +37.25 | [30.25, 44.75] | [27.98, 46.25] |
| Qwen W3 | B | M1−K | +61.25 | [46.50, 75.50] | [43.20, 77.38] |
| Qwen W3 | B | M1−M4 | -6.75 | [−9.51, −4.00] | [−12.01, −2.56] |
| Qwen W3 | S | M1−M0 | +37.50 | [30.50, 45.00] | [28.15, 46.49] |
| Qwen W3 | S | M1−K | +60.50 | [45.25, 75.25] | [41.99, 76.96] |
| Qwen W3 | S | M1−M4 | -6.50 | [−9.25, −3.75] | [−11.73, −2.34] |
| Qwen W1 | B | M1−M0 | +28.04 | [23.48, 32.77] | [21.74, 34.15] |
| Qwen W1 | B | M1−K | +47.97 | [40.71, 54.56] | [36.45, 58.85] |
| Qwen W1 | B | M1−M4 | -2.87 | [−5.41, −0.51] | [−6.55, 0.39] |
| Qwen W1 | S | M1−M0 | +28.04 | [23.48, 32.77] | [21.74, 34.15] |
| Qwen W1 | S | M1−K | +48.31 | [41.05, 54.90] | [36.69, 59.32] |
| Qwen W1 | S | M1−M4 | -2.70 | [−5.07, −0.51] | [−6.26, 0.47] |
| Qwen W4 | B | M1−M0 | +36.96 | [32.32, 41.07] | [30.72, 42.47] |
| Qwen W4 | B | M1−K | +56.25 | [47.86, 64.46] | [43.23, 68.82] |
| Qwen W4 | B | M1−M4 | -1.61 | [−4.64, 1.61] | [−6.58, 3.89] |
| Qwen W4 | S | M1−M0 | +36.96 | [32.32, 41.07] | [30.70, 42.55] |
| Qwen W4 | S | M1−K | +56.43 | [48.21, 64.64] | [43.48, 69.00] |
| Qwen W4 | S | M1−M4 | -1.61 | [−4.64, 1.79] | [−6.70, 3.95] |

All differences are percentage points with unadjusted conditional intervals. Restriction weakens endpoint ambiguity as a sole explanation. It does not equalize optimization difficulty or establish a component mechanism.

Table 38: Original histories and endpoint support of ANSWER_UNSEEN.

| Block | $n/h/c$ | P cases | S cases | Full pairs | Partial pairs |
|---|---|---|---|---|---|
| Qwen W3 | 152/25/57 | 56 | 96 | 25 | 0 |
| Qwen W1 | 106/17/43 | 68 | 38 | 17 | 0 |
| Qwen W4 | 110/19/48 | 62 | 48 | 19 | 0 |
| Llama W3 | 152/25/57 | 56 | 96 | 25 | 0 |
| Phi R1 | 152/25/57 | 56 | 96 | 25 | 0 |
| Phi R2 | 156/26/57 | 66 | 90 | 26 | 0 |

The same membership applies to every compared model and both paths. It contains no Q/R cases. The supplementary exports retain exact case IDs, membership digests and target-label counts.

Table 39: All-cell component compatibility remains a separate requirement.

| World | Path | Contrast | All cells within 3 pp? |
|---|---|---|---|
| Qwen W3 | B | M1−M0 | No |
| Qwen W3 | B | M2−M1 | No |
| Qwen W3 | B | M1−K | No |
| Qwen W3 | B | M2−K | No |
| Qwen W3 | S | M1−M0 | No |
| Qwen W3 | S | M2−M1 | No |
| Qwen W3 | S | M1−K | No |
| Qwen W3 | S | M2−K | No |
| Qwen W1 | B | M1−M0 | No |
| Qwen W1 | B | M2−M1 | No |
| Qwen W1 | B | M1−K | No |
| Qwen W1 | B | M2−K | No |
| Qwen W1 | S | M1−M0 | No |
| Qwen W1 | S | M2−M1 | No |
| Qwen W1 | S | M1−K | No |
| Qwen W1 | S | M2−K | No |
| Qwen W4 | B | M1−M0 | No |
| Qwen W4 | B | M2−M1 | No |
| Qwen W4 | B | M1−K | No |
| Qwen W4 | B | M2−K | No |
| Qwen W4 | S | M1−M0 | No |
| Qwen W4 | S | M2−M1 | No |
| Qwen W4 | S | M1−K | No |
| Qwen W4 | S | M2−K | No |

No means that the conjunction of two-sided interval compatibility conditions is unestablished. It does not imply harm in every cell. Improvement or imprecision can also place an interval outside the margin. The complete per-cell flags and paired effects are retained in machine-readable exports.

### H.5 SOURCE-BOUND RELEASE AND REPRODUCTION

The scientific supplement contains `original_experiments/`, `graded_assignment/`, and `presentation/` modules for the original comparison evidence, graded-assignment extension, and paper displays. It contains protocol and amendment records, source/checkpoint identities, exact design and membership records, six training receipts/traces, qualification evidence, both recorded prediction paths, full estimates and figure/table inputs. Original private locations are replaced only in public copies, with original source hashes distinguished from sanitized-file hashes and a private relocation map retained by the authors.

The retained observation scope includes every row consumed by both complete analyses: 39 unique prediction banks with 673,152 records exposed through 48 condition/path views. Historical rows outside that scope and model weights remain explicitly external. Invalid records and probability/logit fields are retained. Compression and physical deduplication do not select by outcome. Source manifests, membership hashes and exact producer identities distinguish these transformations from fresh inference.

The supplement README separates display regeneration, the historical bounded recorded-data analyses, both complete graded-assignment statistical analyses, and neural execution requiring external weights. Statistical commands run in a fresh output directory outside the release with no model download or GPU execution. Both paths compare all scientific JSON fields and source-table rows under their explicit provenance normalization and inherited $10^{-12}$ floating tolerance. Source-row pointers bind each new paper display and claim to canonical full-precision estimates. Original failures and post-inspection amendments remain disclosed; successful CPU reproduction is not a new neural replication, validation of interval coverage or restoration of the original numerical gates.