OUT OF SIGHT, STILL IN MIND? SEPARATING INTERNAL AND ANSWER-LEVEL EFFECTS OF RETAINED KEY–VALUE HISTORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Cache eviction, prompt compression, and context management change which source tokens remain directly accessible in the key–value (K/V) cache, while later states computed from those tokens can remain. How removed text then influences computation, and whether internal effects translate into answer-level effects, is a measurement question. We study it with a paired cache-interchange design. Two histories specify the same task, by worked examples or by a direct instruction, and are followed by an identical bridge text; we replace the source-prefix cache with a shared neutral cache and swap the retained bridge states between histories under an identical query. We read donor projections at three levels, an intermediate hidden state, the next-token logits, and the correct-minus-incorrect answer margin, with endpoint-specific aggregation, and we report candidate choices wherever candidate scores were archived. Across five arms (Qwen, Mistral, and OLMo; two synthetic rule tasks; 500 indirect-object-identification (IOI) episodes), retained bridge states remain causally active after source-prefix neutralization: mean net donor-axis projections at the intermediate readouts range from 0.59 to 0.97, all 1,780 episodes are positive, and the format difference vanishes when bridge attention is cut. These internal effects do not establish corresponding answer-level effects. Logit-contrast projections are smaller (0.007–0.26), and answer-margin projections range from $+0.42$ to -0.09. On IOI, which the model solves without the histories, a hidden-state projection near one coexists with a small negative donor-oriented margin shift (-0.032 [-0.044, -0.020]) and no changed candidate choice in either patch direction. We formalize what donor projections can and cannot establish and give reporting practices that keep internal alignment, answer margins, and candidate choices separate.

1 INTRODUCTION

Long-context inference increasingly manages memory by changing what the key–value (K/V) cache contains. Eviction policies discard entries judged less useful (Zhang et al., 2023; Li et al., 2024), streaming methods keep a recent window plus a few initial tokens (Xiao et al., 2024), and learned compression replaces text with gist tokens, memory slots, or summaries (Mu et al., 2023; Ge et al., 2024; Kontonis et al., 2026). These methods change which source entries remain directly accessible. Their efficiency objectives do not by themselves determine how removed text continues to influence computation: states cached before removal were computed while the source was visible, so they can retain its influence. MEMENTO reports such an implicit channel in trained compression models (Kontonis et al., 2026), so persistence itself is established. What remains open is how to measure the influence of removed text precisely: at which level of computation it acts, and whether an internal effect implies an effect on the model's answer.

Answering that question requires care, because internal and behavioral evidence can diverge. Information can be decodable from a representation without being used (Hewitt & Liang, 2019; Elazar et al., 2021), interventions can produce clean internal effects through pathways that do not drive behavior (Makelov et al., 2024), and conclusions from activation patching depend on the chosen metric and corruption (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). Two ambiguities are acute for cache interventions. The first is the *level* of measurement: a hidden state read shortly after

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 intervention can move substantially along a contrast axis while the logits and the answer margin move little. The second is *competence*: an effect measured in a model that cannot use the stated rule says little about how retained history shapes competent behavior. Reporting one level at a time, without a competence check, leaves these ambiguities entangled.

We separate them with paired histories that specify the same task, either by worked examples (E) or by a direct instruction (D). Each history is followed by an identical bridge text; we then replace the source-prefix cache with a shared neutral cache and swap the retained bridge states between histories under an identical query (Figure 1). We measure the effects at hidden-state, logit, and answer-margin readouts, together with candidate choices and competence where candidate scores were archived.

The design lets us ask three questions in order. **RQ1 (persistence):** after source-prefix neutralization, do retained bridge states still causally affect later computation? **RQ2 (level):** does a donor-axis effect at an intermediate hidden state accompany corresponding effects on the logit contrast, the answer margin, and candidate choices? **RQ3 (validity):** does interpreting an effect as use of task information depend on whether the model performs the task? We study five arms: Qwen2.5-3B, Mistral-7B, and OLMo-2-7B on synthetic rule tasks, and Qwen on 500 indirect-object-identification (IOI) episodes (Wang et al., 2023), a natural-language task the model solves reliably.

The answers are consistent across arms. For RQ1, retained bridge states remain causally active in the tested construction: mean net donor-axis projections at the intermediate readouts range from 0.59 to 0.97, all 1,780 episodes are positive, and the format difference vanishes under the bridge cut. For RQ2, these internal effects do not establish corresponding answer-level effects: logit-contrast projections are smaller (0.007–0.26), and answer-margin projections range from $+0.42$ to -0.09. On IOI, a hidden-state projection near one coexists with a small negative donor-oriented margin shift and no changed candidate choice in either patch direction. For RQ3, competence changes what a positive projection means: the primary synthetic arm shows positive internal projections while preferring the correct answer in only 50% of episodes. Our contributions follow:

- **Design.** A paired, source-neutralized cache interchange with same-format controls and a complete path cut, defining a causal contrast on retained descendants of removed text (Section 3).
- **Endpoint separation.** A common donor-projection framework applied to hidden states, logits, and answer margins with explicit, endpoint-specific aggregation, together with competence and candidate-choice reporting and formal limits: projection omits orthogonal change, raw logits carry a softmax gauge, and a complete cut yields zero by construction (Section 3.3, Section F).
- **Evidence.** Five arms across three model families showing that retained bridge states are consulted in every case, while logit and answer-margin effects are smaller, task-dependent in sign, and not consistently larger where hidden-state projections are larger (Section 4).
- **Practice.** Reporting rules for cache interventions, retrospective robustness checks, and archived episode records with CPU reanalysis code (Section 5, Sections D and G).

2 Related work

Changing what the cache contains. Efficiency methods remove or compress cached context at inference time. H2O and SnapKV evict entries by attention statistics while preserving task performance (Zhang et al., 2023; Li et al., 2024); StreamingLLM keeps initial "sink" tokens and a recent window (Xiao et al., 2024); gist tokens and in-context autoencoders learn compressed representations of prompts (Mu et al., 2023; Ge et al., 2024). MEMENTO trains models to summarize and mask reasoning blocks and finds that information survives in K/V states computed while a block was visible (Kontonis et al., 2026). We do not evaluate these methods. We study a controlled source-neutralized construction in frozen models with an untrained bridge and ask how retained influence appears at different levels of computation.

Causal interventions and their pitfalls. Causal mediation and interchange interventions replace internal states to test their role in behavior (Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022). Their conclusions depend on corruption method and metric (Zhang & Nanda, 2024); continuous logit-difference metrics are recommended over discrete accuracy (Heimersheim & Nanda, 2024); and an intervention can reproduce an intended output through a dormant pathway unrelated to the mechanism of interest (Makelov et al., 2024). Probing work draws the same line from the other side:

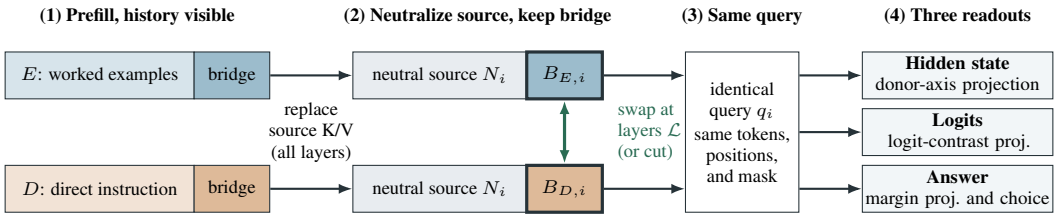

Figure 1: Design. (1) Each history and the identical bridge are prefilled while the history is visible. (2) The source-prefix K/V is replaced at every layer by a shared neutral cache, so the two formats differ only through their retained bridge states (thick borders), which are then swapped at layers $\mathcal{L}$ or masked at every layer (cut). (3) The same query is evaluated on every cache. (4) Donor projections are read at three levels with endpoint-specific aggregation, and candidate choices are recorded where candidate scores were archived.

decodability does not imply use (Hewitt & Liang, 2019; Elazar et al., 2021). We apply these lessons to cache interventions. Our cache-specific contribution combines three elements: a source-neutralized interchange with a deterministic path check, the same intervention read at hidden-state, logit, and answer-margin levels within each episode, and candidate-choice and competence records that show when those readings diverge.

In-context information and format sensitivity. Prompts supply task information without weight updates (Brown et al., 2020), and forward passes can implement learning-like algorithms in controlled settings (Akyürek et al., 2023; Von Oswald et al., 2023). Task and function vectors show that compact internal states can transfer task behavior across contexts (Hendel et al., 2023; Todd et al., 2024). At the same time, demonstrations help partly through their format and label space rather than their input–label correctness (Min et al., 2022), and model behavior is sensitive to prompt formatting, answer position, and option order (Zhao et al., 2021; Sclar et al., 2024; Pezeshkpour & Hruschka, 2024). Examples and direct instructions differ in exactly these surface properties. We therefore treat a donor-directed effect as evidence that retained presentation state is consulted, not as evidence of transferred task knowledge.

3 Design and Estimands

3.1 Paired histories, source neutralization, and interchange

Each episode i pairs an examples history (E) with a direct-instruction history (D) that specify the same task, followed by the identical bridge. The SAME/DIFFERENT task returns an input bit or its complement; DAX/WUG maps two categories to two labels under counterbalanced mappings; IOI asks for the recipient in sentences such as *Laura and Martin went to the store. Martin gave a drink to*, whose answer is Laura. The later query q_i is identical across all cache conditions of an episode. The synthetic queries restate the full mapping; IOI queries are incomplete sentences that the model can already solve. Exact strings are in Sections A and H.

Let $S = \{0, \ldots, H-1\}$ be padded source positions and $B = \{H, \ldots, H+b-1\}$ bridge positions. After processing history $r \in \{E, D\}$ and the bridge, the neutralized cache at layer ℓ is

$$C^{\ell}_{r,i} = \left[N^{\ell}_i[S]\,;\, B^{\ell}_{r,i}[B]\right], \tag{1}$$

where each block holds keys and values and N_i is computed from a neutral history shared by both formats. Replacing the source block removes direct access to the source while preserving whatever the bridge absorbed from it. For recipient r, donor d, and selected layers $\mathcal{L}$, interchange replaces only bridge entries:

$$I_{\mathcal{L}}(C_r, C_d)^{\ell}[t] = \begin{cases} C^{\ell}_d[t], & \ell \in \mathcal{L},\ t \in B, \\ C^{\ell}_r[t], & \text{otherwise.} \end{cases} \tag{2}$$

Recipient source entries, positions, unselected layers, and the attention mask are therefore unchanged. Qwen and OLMo swap layers 0–5 and Mistral layers 0–4; configurations per arm are in Table 3 (Section B.2).

3.2 CONTROLS AND THE IDENTIFIED CONTRAST

Each episode evaluates both cross-format directions ($E \leftarrow D$, $D \leftarrow E$) and two same-format controls, in which a history of the same format but a different session identifier donates its bridge. For fixed parameters, query, positions, and mask, write $f_v(C, q_i)$ for output v under cache C. The effect of one replacement is

$$\tau^v_{i,r\leftarrow d} = f_v(I_{\mathcal{L}}(C_r, C_d), q_i) - f_v(C_r, q_i). \tag{3}$$

Because both terms run with the same recipient configuration, Eq. (3) is a causal effect of the specified state replacement. It is narrower than the total effect of the original prompt: the hybrid cache need not arise in an ordinary forward pass, and the formats differ in wording, length, and repetition as well as in how they convey the task. The contrast therefore identifies influence of retained presentation state, not of an abstract rule variable (Section F).

3.3 DONOR PROJECTIONS AT THREE READOUT LEVELS

For recipient baseline a, opposite-format baseline b, and patched output z, the donor projection is

$$P(a, z, b) = \frac{(z-a)^\top (b-a)}{\|b-a\|_2^2}, \qquad \|b-a\|_2^2 > 10^{-12}, \tag{4}$$

the coordinate of the patched output along the episode's recipient–donor difference axis: $P = 1$ reaches the donor's coordinate on that axis, and movement orthogonal to it is not measured (Proposition 1). Averaging both directions and subtracting the same-format control gives

$$C_i = \tfrac{1}{2}\{P(Y_E, Y_{E\leftarrow D}, Y_D) + P(Y_D, Y_{D\leftarrow E}, Y_E)\}, \tag{5}$$

$$W_i = \tfrac{1}{2}\{P(Y_E, Y_{E\leftarrow E'}, Y_D) + P(Y_D, Y_{D\leftarrow D'}, Y_E)\}, \tag{6}$$

$$\Delta_i = C_i - W_i, \qquad \widehat{\Delta} = \tfrac{1}{n}\textstyle\sum_i \Delta_i. \tag{7}$$

We use this common donor-projection framework with endpoint-specific aggregation. At the *hidden-state* readout, Y is the final-query hidden state (tuple index 8 for Qwen and OLMo, 7 for Mistral), and $\widehat{\Delta}^h$ is the unweighted episode mean of net projections. At the *logit* readout, Y is the next-token logit vector, and $\widehat{\Delta}^\ell$ is aggregated the same way; for IOI we use mean-centered logits $\widetilde{\ell} = \ell - V^{-1}(\mathbf{1}^\top \ell)\mathbf{1}$, which remove additive offsets that leave the softmax unchanged (Proposition 2). Neither logit projection measures full-vocabulary probability change. At the *answer-margin* readout, Y is the scalar margin m between the correct and incorrect candidate, with format gap $g_i = m_{i,D} - m_{i,E}$ and bidirectional scalar projection C_i^m. Because per-episode ratios are unstable when $|g_i|$ is small, we aggregate with absolute-gap weights,

$$\widehat{\Pi}^m_{\times} = \frac{\sum_i |g_i|\, C_i^m}{\sum_i |g_i|} = \frac{\sum_i U_i^m}{\sum_i |g_i|}, \qquad \widehat{\Pi}^m_{\text{net}} = \frac{\sum_i U_i^{m,\text{net}}}{\sum_i |g_i|}, \tag{8}$$

where U_i^m is the archived donor-oriented margin movement (Eq. (12)) and $U_i^{m,\text{net}}$ subtracts the same-format movement (Eq. (20)). Same-format margins were archived only for IOI, so the synthetic arms report the cross-only $\widehat{\Pi}^m_{\times}$ and IOI reports $\widehat{\Pi}^m_{\text{net}}$ (derivation and weighting sensitivity in Section D). A normalized projection hides its scale, so we report the mean gap $\overline{|g|}$ beside it. Finally, the *competence gate* $A_r = n^{-1}\sum_i \mathbf{1}[m_{i,r} > 0]$ records how often the unpatched format prefers the correct candidate, and a *candidate-choice change* occurs when patching flips the sign of the margin. Ties earn no credit, and scores compare two candidates rather than free generation.

3.4 ARMS, PATH CHECK, AND STATISTICS

Three frozen confirmatory arms (Qwen SAME/DIFFERENT, n=512; Mistral SAME/DIFFERENT, n=256; Qwen DAX/WUG, n=256) share one runner. Two later extensions change the apparatus: OLMo SAME/DIFFERENT (n=256) uses different history templates, archives no candidate scores, and uses an unstratified bootstrap; Qwen IOI (n=500, five evaluation templates excluded from source demonstrations, balanced name order) keeps neutral-source padding masked and adds centered logits (Section H). The complete bridge cut masks every bridge position at every layer; with all other accessible state shared, zero difference follows deterministically (Proposition 3), so the cut

Algorithm 1 Paired source-neutralized cache interchange with three readouts (one episode)

This schematic summarizes the shared intervention. Donor construction and archived readouts vary by arm: IOI examples controls also resample source names; OLMo does not enforce $s \neq s'$ and archives no candidate scores (Table 3).

Input: frozen model f; histories $E(s), D(s), E(s'), D(s')$ with $s \neq s'$ (not enforced for OLMo); query q_i; layers $\mathcal{L}$; readout j

Output: $\Delta_i^h, \Delta_i^\ell$; where recorded, margin movement U_i^m, gap g_i, and candidate choices; cut distances

function NEUTRALIZE(x, s) ▷ *Eq.* (1)
  $K \leftarrow$ PREFILL($x \,\|\, bridge$); $N \leftarrow$ PREFILL($neutral(s) \,\|\, bridge$) ▷ *filler masked*
  return $[\, N[S]\,;\, K[B]\,]$ at every layer
$C_E \leftarrow$ NEUTRALIZE($E(s), s$); $C_D \leftarrow$ NEUTRALIZE($D(s), s$) ▷ *shared neutral block*
$C_{E'} \leftarrow$ NEUTRALIZE($E(s'), s'$); $C_{D'} \leftarrow$ NEUTRALIZE($D(s'), s'$)
$Y_E \leftarrow f(q_i \mid C_E)$; $Y_D \leftarrow f(q_i \mid C_D)$ ▷ *baselines*
for each $(r, d) \in \{(E, D), (D, E), (E, E'), (D, D')\}$ **do** ▷ *cross, then same-format*
  $Y_{r \leftarrow d} \leftarrow f\big(q_i \mid I_{\mathcal{L}}(C_r, C_d)\big)$ ▷ *Eq.* (2)*, cloned cache*
$Y_E^{\text{cut}}, Y_D^{\text{cut}} \leftarrow$ baselines with every bridge position masked at every layer
from each Y, read hidden state $h^{(j)}$, logits ℓ (centered where recorded), and margin m (where recorded)
return $\Delta_i^h, \Delta_i^\ell$ by Eq. (7); where recorded, U_i^m, g_i for Eq. (8) and choices $\mathbf{1}[m > 0]$; cut distances

Table 1: Main results at three readouts. *Hidden*: mean net donor-axis projection of the final-query hidden state. *Logit*: mean net projection along the logit contrast (raw logits for archived arms; *mean-centered for IOI, whose raw value is .004 [.002, .007]). *Margin*: absolute-gap-weighted projection of the correct-minus-incorrect margin (Eq. (8)), retrospective; ×cross-format only ($\widehat{\Pi}_{\times}^{m}$) because same-format margins were not archived; IOI is same-format-adjusted ($\widehat{\Pi}_{\text{net}}^{m}$; cross-only value $-.032$). $\overline{|g|}$: mean format gap in the margin (logits). Acc.: strict correct-candidate preference of the unpatched neutralized E/D baselines (%). Brackets: 95% bootstrap intervals, stratified by task cell except OLMo (archived unstratified). SD: SAME/DIFFERENT; DW: DAX/WUG; n/r: not recorded. Hidden-state net projections are positive in every episode of every arm.

Arm	n	Acc. E/D	Hidden $\widehat{\Delta}^h$	Logit $\widehat{\Delta}^\ell$	Margin $\widehat{\Pi}^m$	$\overline{\lvert g \rvert}$
Qwen SD	512	50.0/50.0	.687 [.687, .688]	.258 [.256, .261]	.422× [.411, .434]	1.069
Mistral SD	256	60.5/63.3	.801 [.800, .802]	.059 [.057, .061]	.259× [.241, .278]	0.045
Qwen DW	256	52.3/61.7	.605 [.604, .607]	.047 [.043, .051]	−.091× [−.113, −.068]	0.538
OLMo SD	256	n/r	.587 [.584, .590]	.033 [−.038, .107]	n/r	n/r
Qwen IOI	500	100/100	.968 [.967, .970]	.007* [.005, .010]	−.032 [−.044, −.020]	0.552

verifies path isolation rather than revealing a mechanism. The paired episode is the unit of analysis. Intervals are percentile bootstraps stratified by task cell (10,000 draws; 100,000 for IOI; OLMo as archived; 20,000 for answer-margin projections), conditional on each arm's fixed templates and layers (Sections C and D). Answer-margin projections, competence rates, directional margin summaries, and leave-one-template-out checks are retrospective analyses of archived records and are labeled as such. Algorithm 1 states the procedure.

4 RESULTS

We answer the three research questions in order. Table 1 carries the central evidence: each row is one arm, and its columns report the same intervention at three readouts, each with its own aggregation.

4.1 RQ1: RETAINED BRIDGE STATES REMAIN CAUSALLY ACTIVE

Source-prefix neutralization does not remove the formats' influence in the tested construction. In every arm, swapping early-layer bridge states produces a large positive donor-axis projection at the intermediate hidden-state readout: means range from 0.587 (OLMo) to 0.968 (IOI), all intervals exclude zero, and all 1,780 episodes are positive (Table 1, Hidden). The same-format control is small throughout, between 0.001 and 0.004 (Table 12 and Section H.2), so replacing a bridge with one from the same format barely moves the state along the cross-format axis. When every bridge position

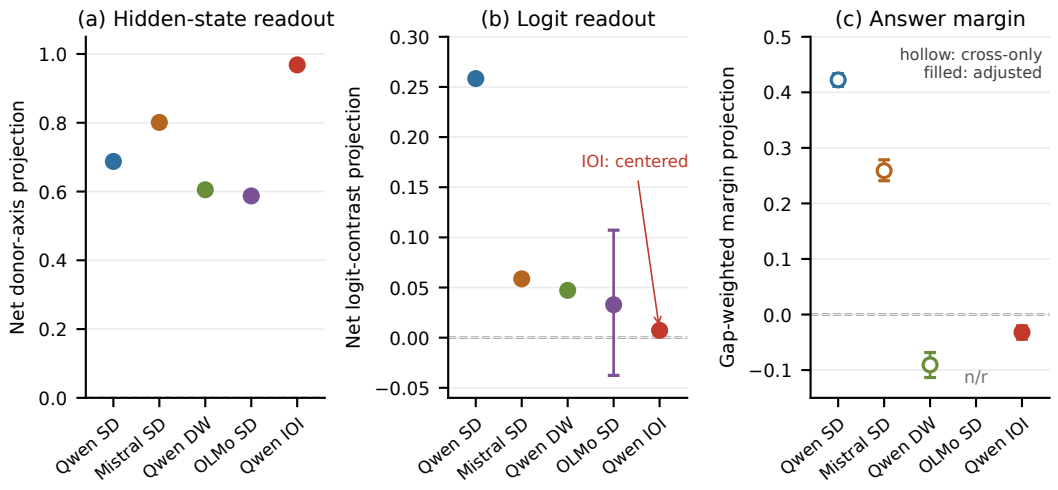

Figure 2: Donor projections at three readouts, shown in separate panels because each uses its own space and aggregation (values and 95% intervals from Table 1; many intervals are narrower than the markers). Hidden and logit panels show unweighted means of net episode projections. The margin panel shows absolute-gap-weighted projections: hollow markers are cross-format only (Qwen SD, Mistral SD, Qwen DW) and the filled marker is same-format-adjusted (IOI). OLMo has no archived margins.

is masked, the baseline difference between formats falls to exactly zero in all arms and episodes, as Proposition 3 requires of a correct implementation. Together these results locate the surviving influence in the retained bridge: after the source is replaced, what the bridge absorbed from it still shapes later computation. The next question is whether that internal effect reaches the answer.

4.2 RQ2: Internal effects do not establish answer-level effects

The three readouts give different pictures of the same intervention (Table 1 and Figure 2). Logit-contrast projections are smaller than hidden-state projections in every arm: 0.258 for primary Qwen, 0.047–0.059 for Mistral and DAX/WUG, 0.033 with an interval spanning zero for OLMo, and 0.007 (centered) for IOI. Answer-margin projections differ in sign across tasks: $+0.42$ for primary Qwen and $+0.26$ for Mistral (both cross-only), -0.09 for DAX/WUG, and -0.032 for IOI, whose interval excludes zero. Larger hidden-state projections do not consistently accompany larger or more positive margin projections in these arms. IOI has the largest hidden-state projection and a small negative donor-oriented margin projection, while primary Qwen has a moderate hidden-state projection and the largest margin projection. Retrospective episode-level correlations point the same way for the hidden-state readout: in every arm, the rank correlation between hidden-state projection and donor-oriented margin movement is negative or near zero, for both the pooled and the aggregate within-cell estimates (-0.36 to $+0.04$; Table 7). Logit-contrast correlations are positive when pooled in the synthetic arms but vary in sign within cells, so we do not read them as consistent tracking. We report the three readouts side by side rather than as ratios, because they use different spaces, normalizations, and aggregation rules.

The directional IOI records show what a small negative margin projection means in practice (Table 2). Each patch direction changes the recorded margin in most episodes (476 and 473 of 500), by 0.064–0.068 logits on average against baseline margins of about 6.2–6.6 logits. Oriented episode by episode toward the donor's baseline, both directions move slightly away from it on average (-0.016 and -0.019 logits), with only 211 and 189 of 500 episodes moving toward the donor. Neither direction changes the preferred candidate in any episode, and the mean absolute change in two-candidate probability is below 0.03 percentage points. The intervention therefore changes margins while preserving every candidate choice. Scale matters for the synthetic arms as well: Mistral's cross-only projection of 0.26 applies to a mean gap of 0.045 logits, about 0.01 logits of movement, and primary

Table 2: IOI answer margins by patch direction (retrospective, $n = 500$ episodes). Change is patched minus recipient-baseline margin (logits); *toward donor* orients each episode's change by the sign of that episode's format gap, so positive values move toward the donor's baseline (same-format rows use the cross-format axis). The donor lies on average at -0.457 logits relative to the examples baseline, so a positive raw change for $E \leftarrow D$ is consistent with movement away from the donor. Choice changes count sign flips of the margin. $|\Delta p|$ is the mean absolute change in probability normalized over the two candidates (percentage points), not a full-vocabulary quantity.

Patch	Mean change	Toward donor	Mean |change|	Changed	Choice changes	$\|\Delta p\|$ (pp)
$E \leftarrow D$ (cross)	$+0.0292$	-0.0163	0.0683	476/500	0/500	0.021
$D \leftarrow E$ (cross)	-0.0329	-0.0187	0.0635	473/500	0/500	0.028
$E \leftarrow E'$ (same)	$+0.0006$	-0.0003	0.0118	385/500	0/500	0.005
$D \leftarrow D'$ (same)	$+0.0008$	$+0.0009$	0.0101	365/500	0/500	0.006

Qwen's 0.42 applies to a 1.07-logit gap between formats that make identical unpatched choices in every episode. Patched candidate choices were not archived for the synthetic arms.

The logit endpoint also varies across fixed conditions. In IOI, the centered-logit projection is positive in 304 of 500 episodes, and two of ten template-by-order cells have negative means (Figure 4b). Separately, OLMo illustrates the same heterogeneity: its hidden-state projection is positive in all 256 episodes, but its logit projection is positive in only 119, and four fixed task cells have means from -0.357 to $+0.660$ that partly cancel, with between-cell differences accounting for 43.9% of the observed variance (Table 10 and Figure 7). Aggregate logit effects can therefore conceal systematic condition-level structure, which is why we report cells alongside means.

These readings are robust to reasonable analysis choices but also bounded by the design. The signs of all three aggregate IOI endpoints persist when any single evaluation template is excluded, which shows that no included template drives the aggregate direction without establishing generalization to unseen templates; margin magnitudes depend on aggregation (unweighted means 0.48, 0.14, -0.15, and -0.07), but their signs do not (Section D.1). The hidden-state readout sits two blocks after the last patched layer, so six of the eight attention layers the query passes through before it (five of seven for Mistral) read donor bridge states, whereas the logits and margin are computed after all layers, most of which read the recipient's own bridge. Readout proximity may therefore account for part of the hidden-state magnitude; it is the most consequential untested alternative, and Section I specifies a depth-control experiment. IOI's positive logit projection persists after centering, and both its raw and centered values are numerically small (.004 and .007), so a common logit offset (Proposition 2) does not explain it.

4.3 RQ3: Competence changes what a positive projection means

The synthetic arms show that positive internal projections do not by themselves establish reliable use of task information. Primary Qwen prefers the correct candidate in exactly 50.0% of episodes under both formats (Figure 3): it fails every SAME case and succeeds in every DIFFERENT case, choosing the answer named in the query's second mapping clause in all 512 episodes. Mistral (60.5/63.3%) and Dax/Wug (52.3/61.7%) show similarly structured, cell-dependent failures. Such fixed preferences resemble documented recency and position biases (Zhao et al., 2021; Pezeshkpour & Hruschka, 2024), but the favored position differs across arms, so we treat the alignment as descriptive rather than causal (Section E). Because the same internal signature appears in arms with substantial competence failures, a hidden-state projection alone cannot distinguish a task-relevant trace from a presentation marker (Eq. (18)).

IOI supplies the complementary case, a task the model performs reliably. Query-only candidate accuracy is 499/500, and both neutralized formats and each patch direction are correct in 500/500 episodes, in both name orders (Figure 4a). The model's competence therefore does not depend on the histories, and the experiment does not establish that either history teaches IOI. IOI therefore shows that the readout dissociation of Section 4.2 also occurs during reliable candidate discrimination.

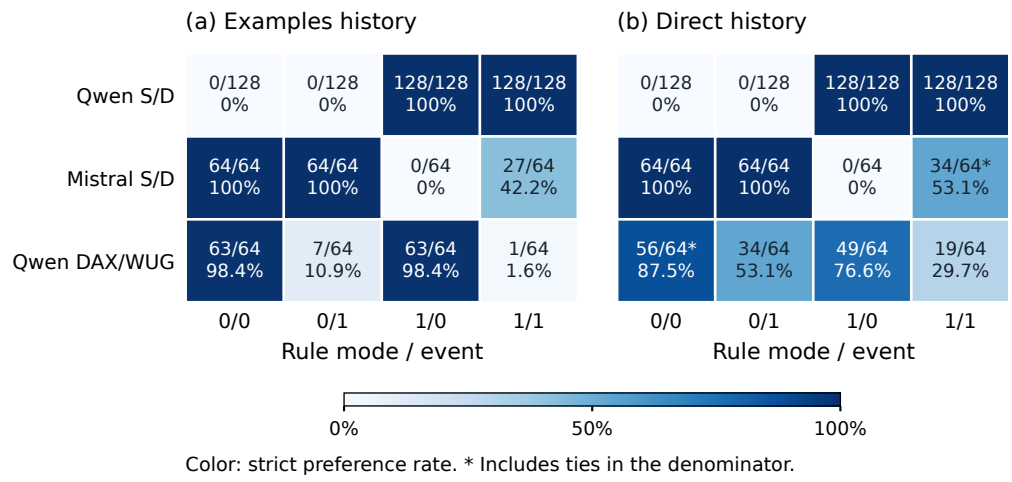

Figure 3: Competence gate for the synthetic arms: strict correct-candidate preference of the unpatched neutralized baselines by task cell. Each entry is the count with margin > 0 over the cell size; ties receive no credit. For SAME/DIFFERENT, c is the rule and x the input bit; for DAX/WUG, the mapping and category. Counts are in Table 9.

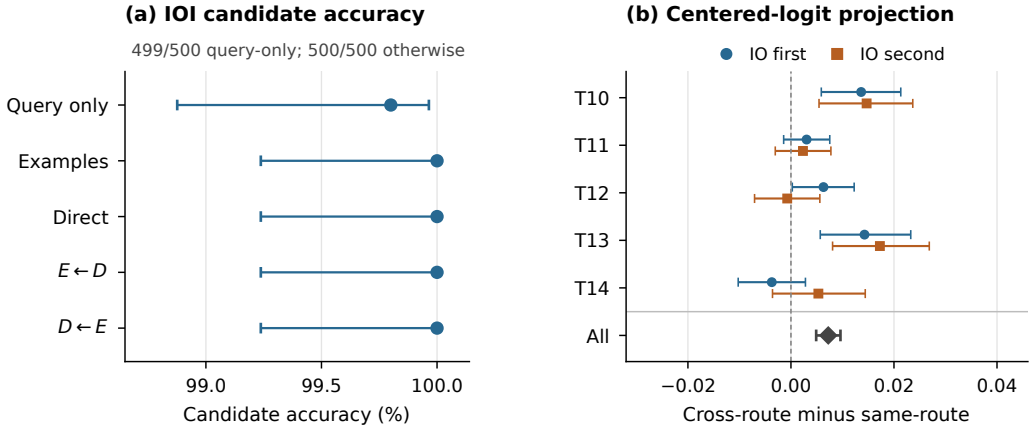

Figure 4: IOI competence and logit heterogeneity. (a) Strict preference for the correct name over the repeated subject; Wilson 95% intervals with $n = 500$ per condition (the two patch directions share episodes and are not pooled). $E \leftarrow D$: examples recipient, direct donor. (b) Net centered-logit projection by evaluation template and name order ($n = 50$ per cell); exploratory within-cell bootstrap intervals; the diamond is the aggregate estimate with its primary interval. Intervals are conditional on the five fixed templates.

5 DISCUSSION, PRACTICE, AND LIMITATIONS

A single normalized internal projection is insufficient to characterize the consequences of a cache intervention. An internal effect remains a valid causal effect of the intervention, but it does not by itself establish an answer-level effect or the use of task information. We evaluate no eviction or compression policy, so the results do not establish effects of deployed cache deletion. They motivate a testable audit practice: measure retained influence at answer-level endpoints, and establish competence before interpreting an intervention as evidence of task-information use.

Reporting practice. These findings translate into five rules for cache interventions. (1) Check competence before interpreting an intervention as evidence of task-information use or transfer, using candidate scores and, where possible, generated answers. (2) Report hidden-state, logit, and answer-margin projections separately, each with its aggregation rule, same-format adjustment status, and denominator, together with candidate-choice changes. (3) Use gauge-free logit measures such as centered logits, full-distribution distances where vectors are retained, and continuous answer measures such as margins (Heimersheim & Nanda, 2024). (4) Treat complete cuts and identity swaps as implementation checks, not as findings. (5) Match donor controls to the contrast of interest; our same-format control changes only a session identifier (and, in IOI examples, source names), which is weaker than a format change. Section I specifies the corresponding experiments.

Specificity. Examples and direct instructions differ in length, repetition, vocabulary, and syntax, so the retained difference can encode any of these. Under the mixture construction of Eq. (18), a format marker alone can yield a positive net projection even when it carries no task-relevant information; a factorial design crossing presentation format with task content, matched for length and repetition, is needed to attribute the effect to how a task was presented rather than to what was written.

Measurement. Projection ignores orthogonal displacement (Proposition 1), and the archived records keep only bidirectional averages at the hidden-state and logit readouts, so direction-specific and full-distribution analyses are not possible retrospectively. Readout proximity may inflate the hidden-state projection. Candidate scores are two-candidate comparisons at one token; in SAME/DIFFERENT, candidates are appended directly after `Answer:` without a separating space, which may contribute to the competence failure and is untested.

Scope and provenance. Each arm uses one evaluation template family and fixed layers, so intervals describe episode variation within those choices, not generalization across tasks or models. IOI changes the padding mask, prompts, and donor construction together, so its differences from the synthetic arms cannot be attributed to the task alone. The frozen evaluation template and the IOI templates were seen during development; the IOI protocol states its criteria were fixed before the final run, but its archived provenance redacts the protocol hash, so that chronology cannot be independently verified (Section H.3). Answer-margin projections, competence analyses, and robustness checks are retrospective.

6 CONCLUSION

After source-prefix neutralization, retained bridge states remain causally active: across five arms and three model families, swapping them produces large donor-axis projections at an intermediate hidden state, and cutting them removes the format difference. That internal effect does not establish corresponding answer-level transfer. Logit-contrast projections are smaller, answer-margin projections differ in sign across tasks, and on competent IOI candidate discrimination a hidden-state projection near one coexists with a small negative donor-oriented margin shift and unchanged candidate choices. Reporting hidden-state, logit, and answer-margin projections separately, with their aggregation rules and a competence check, turns a single suggestive number into a calibrated account of what retained history does. The most informative next steps are a depth-control experiment that varies patch location and readout depth, and a factorial specificity design that separates presentation format from task content.

AI USE STATEMENT

We used large language model tools to assist with experimental design discussions, analysis code and code review, literature search, interpretation, drafting and editing, and figure and LaTeX preparation. All analyses reported in this manuscript were computed on CPU from archived episode records, and no model inference was rerun for it; the scripts in the supplementary archive reproduce every reported statistic. The authors reviewed all citations, numbers, and claims and take full responsibility for the content of this paper.

ETHICS STATEMENT

The experiments use released model weights and constructed prompts, with no human participants or private data. The main interpretive risk is overstating what internal-state interventions establish about task behavior or about deletion of context in deployed systems. We therefore report competence, effects at three readouts, candidate choices, inconclusive outcomes, metric limits, and provenance gaps together.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 define the design and estimands; Sections A, B and H give prompts, masks, scoring, model revisions, and implementation algorithms; Sections C and D define all estimators, bootstrap procedures, and robustness checks; Section F proves the formal qualifications. The supplementary archive contains every episode record and CPU scripts that reproduce all reported statistics and figures without model weights (Section G). Regenerating activations requires the pinned models and a compatible GPU environment.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. 2023. URL https://arxiv.org/abs/2211.15661.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uREj4ZuGJE.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. 2021. URL https://arxiv.org/abs/2106.02997.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Vasilis Kontonis, Yuchen Zeng, Shivam Garg, Lingjiao Chen, Hao Tang, Ziyan Wang, Ahmed Awadallah, Eric Horvitz, John Langford, and Dimitris Papailiopoulos. Memento: Teaching llms to manage their own context, 2026. URL https://arxiv.org/abs/2604.09852.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 22947–22970. Curran Associates, Inc., 2024. doi: 10.52202/079017-0722. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/28ab418242603e0f7323e54185d19bde-Paper-Conference.pdf.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/70b8505ac79e3e131756f793cd80eb8d-Abstract-Conference.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Jesse Mu, Xiang Li, and Noah Goodman. Learning to compress prompts with gist tokens. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 19327–19352. Curran Associates, Inc., 2023. doi: 10.52202/075280-0848. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3d77c6dcc7f143aa2154e7f4d5e22d68-Paper-Conference.pdf.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RIu5lyNXjT.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. 2024. URL https://arxiv.org/abs/2310.15213.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. 2024. URL https://arxiv.org/abs/2309.16042.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang "Atlas" Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34661–34710. Curran Associates, Inc., 2023. doi: 10.52202/075280-1506. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6ceefa7b15572587b78ecfcebb2827f8-Paper-Conference.pdf.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

A EXACT TASK CONSTRUCTION AND ANSWER SCORING

A.1 FROZEN EVALUATION HISTORIES

Both formats are raw tokenized text, without a chat-template wrapper. The evaluated history template is template 3. A session nonce has four independently sampled digits separated by spaces. The distinct same-format donor nonce is redrawn if it equals the recipient nonce. Examples/direct baselines and cross-format donors share the first nonce. The following verbatim strings use `{s}` for that nonce; braces denote substitution, not literal prompt tokens.

```
Session {s}.  Demonstrations:  0 gives 0, 1 gives 1, 0 gives
0, 1 gives 1.
Session {s}.  Stated instruction:  SAME.
```

The DIFFERENT example replaces outputs by their complements and the direct instruction replaces SAME by DIFFERENT. Each source string ends in a newline. The neutral history is

```
Session {s}.  Earlier material was neutral:  cloud stone
window copper river lamp paper chair.
```

It also ends in a newline. Every source is padded to the maximum probed history length before appending the identical bridge string `History processed.  Ready.\n`. The neutral prefix replaces the padded source block at every layer, including originally padded positions.

For DAX/WUG, the example format uses four name/category/output triples: mira/DAX, tovan/WUG, selo/DAX, and pira/WUG. With $a = X, b = Y$ in mapping mode 0 and $a = Y, b = X$ in mode 1, the evaluation formats are

```
Session {s}.  Demonstrations:  mira with DAX gives {a};
tovan with WUG gives {b}; selo with DAX gives {a}; pira with
WUG gives {b}.
Session {s}.  Stated mapping:  answer {a} for DAX and {b} for
WUG.
```

Later names are drawn from mira, tovan, selo, pira, navi, lumo, keta, vori, zema, riko, fena, and daro. Because four names also occur in demonstrations, "new item" need not mean an unseen name. The later query provides the item's category explicitly.

A.2 CURRENT QUERIES

The SAME/DIFFERENT query is

```
Current instruction:  {rule}.  New bit:  {x}.  SAME means
answer {x}.  DIFFERENT means answer {1-x}.  Reply only 0 or
1.
Answer:
```

There is a newline before `Answer:`, with no additional final newline. Both candidate mappings are explicitly stated, including the answer for the current bit. Thus current-text equality is particularly strong, but competent use of that text must still be checked; Section 4.3 shows the primary apparatus has a systematic per-cell preference pattern. Its association with candidate position is descriptive, not an identified mechanism.

The amended DAX/WUG query is

```
Current mapping:  DAX -> {a}; WUG -> {b}.  New item:  {name}
has {category}.  Reply only X or Y.
Answer:\n
```

The final newline was added in the secondary amendment to preserve the query's token boundary when a candidate answer was appended. The first secondary launch stopped on its first episode because that boundary changed. The earlier version-1 launch also failed a bare-label tokenization assumption on Mistral. These are recorded apparatus failures, not scientific null outcomes.

A.3 CONTEXTUAL CANDIDATE SCORES

Let $u = \mathrm{tok}(q)$ and let $v_+ = \mathrm{tok}(q \,\|\, a_+)$ and $v_- = \mathrm{tok}(q \,\|\, a_-)$ for correct/incorrect answer strings. The runner checks that u is an exact prefix of both. Let k be the first position where v_+ and v_- differ after their shared prefix. If necessary, a second forward pass conditions on that shared token prefix. With the resulting scoring logits ℓ, the recorded quantities are

$$m = \ell_{v_+[k]} - \ell_{v_-[k]}, \qquad p_+ = \frac{\exp(\ell_{v_+[k]})}{\exp(\ell_{v_+[k]}) + \exp(\ell_{v_-[k]})} = \sigma(m). \tag{9}$$

This is a two-candidate normalization at one distinguishing token. It is not the probability assigned to the entire answer string under the full vocabulary, and it is not free generation accuracy. The main hidden/logit outcomes are taken at the original final query position; the contextual candidate scorer can use a later position if the answer candidates share a prefix.

B ARCHIVED IMPLEMENTATION, MASKS, AND REPRODUCIBILITY LOCKS

B.1 CACHE OPERATIONS

Per-layer key and value tensors use token positions on dimension 2. Prefix forwards initially mask EOS filler positions. Source neutralization concatenates neutral-prefix keys/values with history-conditioned bridge keys/values for every layer. The intervention copies bridge slices for the selected layer set and clones all other slices. Later evaluation clones the cache again so one condition cannot mutate another through cache extension. Current-query positions begin immediately after the padded source plus bridge.

The original later-query attention mask enables all prefix positions, including the neutral filler positions previously masked. The bridge-cut mask then disables positions B across all layers. Algorithm 2 reproduces this behavior explicitly. The IOI extension instead preserves the neutral-prefix padding mask (Algorithm 4); because its prompts and donor construction also change, it is not an isolated masking ablation.

Algorithm 2 Cache construction and query evaluation as implemented in the frozen arms

Input: model f; tokenizer; padded source length H; bridge tokens β with $|\beta| = b$; layers $\mathcal{L}$; readout j

function PREFILL(x)
  $t \leftarrow \mathrm{tok}(x)$ without special tokens; $\pi \leftarrow$ EOS padding to length H
  $\mu \leftarrow [\mathbf{1}_{|t|};\ \mathbf{0}_{H-|t|};\ \mathbf{1}_b]$ ▷ *filler masked while prefilling*
  return all-layer K/V of $f([t; \pi; \beta]$ with mask $\mu)$
function NEUTRALIZE(K, N)
  for each layer ℓ **do**
    $C^\ell \leftarrow [\, N^\ell[0{:}H)\,;\ K^\ell[H{:}H{+}b)\,]$ ▷ *neutral source, retained bridge*
  return clone of C
function PATCH(C_r, C_d)
  $C \leftarrow$ clone of C_r
  for each layer $\ell \in \mathcal{L}$ **do**
    $C^\ell[H{:}H{+}b) \leftarrow C_d^\ell[H{:}H{+}b)$ ▷ *bridge slice only*
  return C
function EVALUATE(C, q, *cut*)
  positions $\leftarrow H{+}b, \ldots, H{+}b{+}T{-}1$; mask $\leftarrow \mathbf{1}_{H+b+T}$ ▷ *neutral filler becomes visible*
  if *cut* **then**
    mask$[H{:}H{+}b) \leftarrow 0$ ▷ *bridge hidden at every layer*
  run f on q with a clone of C in inference mode
  return hidden state $h^{(j)}$ and raw logits at the final query token; contextual margin m (Section A.3)

B.2 PINNED MODELS AND SOFTWARE

The frozen Qwen revision is `aa8e72537993ba99e69dfaafa59ed015b17504d1` for `Qwen/Qwen2.5-3B-Instruct`. Mistral uses `c170c708c41dac9275d15a8fff4eca08d52bab71` for `mistralai/Mistral-7B-Instruct-v0.3`. OLMo uses `470b1fba1a`

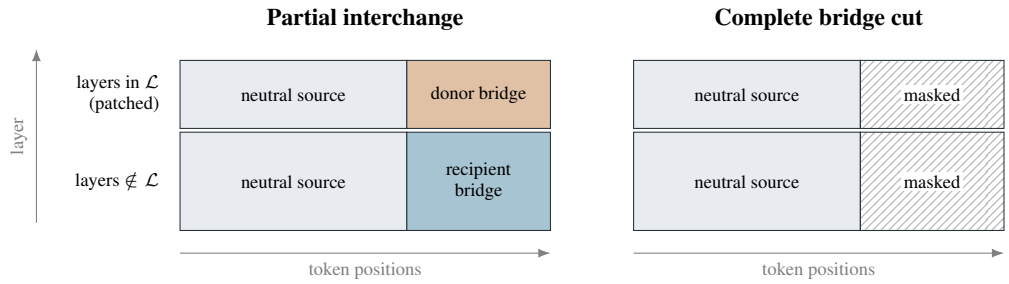

Figure 5: The interchange and the cut are different interventions on the same neutralized cache. Left: interchange replaces only bridge entries in the selected layers $\mathcal{L}$; source entries and all other layers keep the recipient's states. Right: the cut hides every bridge position from the query at every layer. A zero difference under the cut therefore shows that the format difference is confined to the bridge, not that the selected layers are uniquely necessary.

`e01581f270116362ee4aa1b97f4c84` for `allenai/OLMo-2-1124-7B-Instruct`. Model and tokenizer revisions are pinned in the frozen runners.

The recorded runtime uses PyTorch 2.6.0, CUDA 12.4, Transformers 5.13.0, SentencePiece 0.2.1, Accelerate 1.10.1, float16, and SDPA attention. The scripts use CUDA explicitly. A successful import check is not a GPU or model-validity check. Exact reproduction must retain revisions, tokenizer behavior, numerical format, positions, mask construction, and cache API semantics. The default public reference runner omits several frozen endpoints and must not be substituted for the confirmatory runner.

Table 3 summarizes the configuration of every arm. Both frozen packages carry canonical manifest hashes that lock runners, protocols, analysis code, and seeds; the hashes are included in the supplementary archive and omitted from the review copy to preserve anonymity. Hash verification checks that files match their recorded locks. It does not certify when an investigator viewed results. The archived chronology records development, invalid launches, preflight checks, and the secondary amendment, and we preserve that distinction rather than treating all development as prospectively unseen.

Table 3: Arm configurations. Patch layers are zero-based; hidden indices refer to the returned tuple, whose entry 0 is the embedding output. The padding column describes neutral-source filler positions during the later query.

Model	Task	Episodes	Patch layers	Hidden index	Bridge tokens	Padding
Qwen2.5-3B	SAME/DIFFERENT	512	0–5	8	5	Exposed
Mistral-7B	SAME/DIFFERENT	256	0–4	7	7	Exposed
Qwen2.5-3B	DAX/WUG	256	0–5	8	5	Exposed
OLMo-2-7B	SAME/DIFFERENT	256	0–5	8	5	Exposed
Qwen2.5-3B	IOI	500	0–5	8	5	Masked

C STATISTICAL DEFINITIONS AND ANALYSIS PROVENANCE

C.1 FROZEN BOOTSTRAP AND SIGN-FLIP CALCULATION

For cell $g \in \{1, 2, 3, 4\}$ containing n_g observations, the stratified bootstrap independently resamples n_g episode indices with replacement. A replicate is

$$\widehat{\Delta}^{*(b)} = \frac{1}{n} \sum_{g=1}^{4} \sum_{j=1}^{n_g} \Delta_{I_{gj}^{(b)}}, \qquad I_{gj}^{(b)} \sim \mathrm{Uniform}(\mathcal{I}_g). \tag{10}$$

The 2.5th and 97.5th percentiles of 10,000 replicates form the 95% interval. Frozen 90% intervals use the same bootstrap distribution for the secondary equivalence criteria. The sign-flip calculation

draws 100,000 vectors of independent signs and reports

$$p_{\text{flip}} = \frac{1 + \sum_{b=1}^{100000} \mathbf{1}\left[\left|n^{-1}\sum_i \epsilon_i^{(b)}\Delta_i\right| \geq |\widehat{\Delta}|\right]}{100001}, \quad \epsilon_i^{(b)} \in \{-1, +1\}. \tag{11}$$

These values require sign-exchangeability under the reference null; a zero mean alone is insufficient. Exchanging examples/direct labels does not negate the symmetric bidirectional statistic, so the calculation is not a design-based randomization test induced by assignment of those labels. They are not uncertainty over the choice of tasks, templates, or models. Fixed layers and evaluation wording were selected during development and retained in the frozen arms. The later IOI primary analysis is specified separately in Section H.2; exploratory diagnostics are labeled as such.

The OLMo extension resamples the complete 256-record list without stratification, using the recorded Python-random seeds and 10,000 draws. Its interval therefore also reflects changing cell proportions in bootstrap samples. The original interval is retained in Table 1; it is not silently replaced with a narrower stratified estimate after observing condition heterogeneity.

C.2 Secondary answer endpoints

Let $d_i^m = m_{i,D} - m_{i,E}$ and use $\text{sgn}(0) = 0$. The frozen donor-oriented margin movement is

$$U_i^m = \frac{\text{sgn}(d_i^m)}{2}\left[(m_{i,E\leftarrow D} - m_{i,E}) - (m_{i,D\leftarrow E} - m_{i,D})\right]. \tag{12}$$

The probability version substitutes the two-candidate probabilities and their own baseline-gap sign. Unlike Eq. (7), these secondary measures do not subtract same-format movement; the asymmetry is inherited from the frozen protocol and is flagged rather than repaired retrospectively, since a repaired version would require records that were not retained. The sign points toward the donor baseline, not toward better task performance. Section D shows that U_i^m is exactly the numerator of the absolute-gap-weighted answer-margin projection used in the main text. The frozen equivalence decision requires both 90% intervals to fit inside $[-0.20, 0.20]$ for the mean margin movement and $[-0.05, 0.05]$ for the mean probability movement. A signed average inside these bounds does not imply small absolute changes for each episode.

C.3 Format-classifier generalization

Reference centroids average hidden states from templates 0–2 with independent reference nonces. For an evaluation example of mapping c and query type x, the classifier compares its hidden state to format centroids from mapping $1 - c$ and the same x:

$$\widehat{r}(h) = \arg\min_{r\in\{E,D\}} \|h - \mu_{r,1-c,x}\|_2. \tag{13}$$

The code assigns ties to the direct format. It averages correctness across the two baseline formats and donor-label agreement across the two patched formats. The following figures concern classifier generalization; the patched donor itself remains matched on c. Template 3 is excluded from centroid

Table 4: Frozen cross-content format-classifier means and archived 95% intervals.

Model	Baseline format classification	Donor-label classification after swap
Qwen	83.50% [82.13, 84.86]	75.59% [75.20, 76.07]
Mistral	91.99% [89.84, 94.14]	79.88% [77.15, 82.42]

estimation but is not investigator-unseen wording. DAX/WUG does not have this centroid analysis in the frozen runner.

D Answer-margin projection

Definition. Applying Eq. (4) to the scalar margin m gives, for recipient E, $P(m_E, m_{E\leftarrow D}, m_D) = (m_{E\leftarrow D} - m_E)/g$ with $g = m_D - m_E$, and for recipient D, $(m_{D\leftarrow E} - m_D)/(-g)$. Their average

is

$$C_i^m = \frac{(m_{i,E\leftarrow D} - m_{i,E}) - (m_{i,D\leftarrow E} - m_{i,D})}{2g_i}, \qquad \text{so} \qquad |g_i|\, C_i^m = U_i^m \tag{14}$$

by Eq. (12). The frozen secondary endpoint is therefore the numerator of a projection whose denominator the frozen analysis did not report. Per-episode ratios C_i^m are unstable when $|g_i|$ is near zero, so the main text aggregates with absolute-gap weights (Eq. (8)), weighting each episode by the size of the margin difference being transferred. This aggregation differs from the unweighted episode means used at the hidden-state and logit readouts, and the synthetic estimate $\widehat{\Pi}_\times^m$ is cross-format only, whereas the IOI estimate $\widehat{\Pi}_{\text{net}}^m$ uses $U_i^{m,\text{net}}$, the cross-format movement minus the same-format movement (Eq. (20)). For episode-level summaries we write $R_i^{m,\times} = U_i^m/|g_i|$ (which equals C_i^m) and $R_i^{m,\text{net}} = U_i^{m,\text{net}}/|g_i|$, both for $g_i \neq 0$. The three readouts therefore share a projection formula but not an aggregate estimand. Episodes with $g_i = 0$ contribute to neither sum; there are 4, 3, 3, and 2 such episodes in the Qwen SAME/DIFFERENT, Mistral, DAX/WUG, and IOI arms.

Estimation. Intervals are 95% percentile intervals of the ratio estimator from 20,000 bootstrap draws. The implementation uses NumPy's `default_rng(20260925)`, created once per arm; each draw visits the task cells in sorted order (four mode-by-event cells for synthetic arms, ten template-by-order cells for IOI) and resamples each cell's episode indices with replacement at the cell's own size. This analysis was defined after the frozen results were known, so it is retrospective; it uses only archived margins and requires no model inference. The frozen primary intervals (10,000 stratified draws) and the IOI primary intervals (100,000 stratified draws, seed 771903) are unchanged. Algorithm 3 gives the computation, and the scripts `decision_level_projection.py` and `robustness_checks.py` in the supplementary archive reproduce every value in Tables 5 to 8 and Table 2 from the saved records.

Table 5: Answer-margin projection by arm (retrospective). $\overline{|g|}$ is the mean absolute format gap in the correct-minus-incorrect margin (logits); "median" is the median of $R_i^{m,\times}$ for the synthetic arms and of $R_i^{m,\text{net}}$ for IOI, over episodes with $g_i \neq 0$; $U > 0$ counts episodes moving toward the donor. Synthetic arms report $\widehat{\Pi}_\times^m$ (cross-format only); IOI reports $\widehat{\Pi}_{\text{net}}^m$, with cross-only value -0.032 $[-0.044, -0.020]$ and same-format value 0.0005 $[-0.0010, 0.0021]$.

Arm	$\overline{\lvert g\rvert}$	Mean U	$\widehat{\Pi}^m$ [95% CI]	Median R^m	$U > 0$
Qwen SAME/DIFFERENT	1.069	.4517	.422 [.411, .434]	.427	374/512
Mistral SAME/DIFFERENT	0.045	.0116	.259 [.241, .278]	.300	142/256
Qwen DAX/WUG	0.538	−.0487	−.091 [−.113, −.068]	−.059	98/256
Qwen IOI (net)	0.552	−.0178	−.032 [−.044, −.020]	−.028	197/500

Interpretation. The answer-margin projection asks how much of the format's effect on the margin travels with the swapped bridge states. It is not candidate accuracy. A large value on a tiny gap, as for Mistral, is a tiny absolute change. For primary Qwen, both unpatched formats make identical choices in every episode, but patched candidate choices were not archived for the synthetic arms, so whether patching changed any synthetic choice cannot be recovered from the records. For IOI, patched choices were archived: neither direction changes any choice (Table 2), and the formats' 0.55-logit margin difference moves slightly away from the donor.

D.1 ROBUSTNESS OF THE ANSWER-LEVEL ANALYSIS

All checks in this subsection are retrospective CPU analyses of archived records.

Aggregation. Table 6 compares the absolute-gap-weighted projection with the unweighted mean and median of per-episode projections over $g_i \neq 0$. Magnitudes depend on the aggregation rule, and signs do not. We keep the gap-weighted estimator in the main text because unweighted means give episodes with near-zero gaps unbounded influence.

Table 6: Sensitivity of the answer-margin projection to aggregation (retrospective). Episode-level summaries use $R_i^{m,\times}$ for the synthetic arms and $R_i^{m,\mathrm{net}}$ for IOI, over episodes with $g_i \neq 0$.

Arm	Unweighted mean R^m	Median R^m	Gap-weighted $\widehat{\Pi}^m$
Qwen SAME/DIFFERENT (cross)	.4823	.4265	.4224
Mistral SAME/DIFFERENT (cross)	.1415	.3000	.2589
Qwen DAX/WUG (cross)	−.1491	−.0588	−.0906
Qwen IOI (net)	−.0662	−.0279	−.0322

Episode-level association. Table 7 reports pooled Spearman and aggregate within-cell rank correlations between episode-level projections and donor-oriented margin movement. Pooled correlations combine variation within fixed task cells with differences between cells. To separate the two, we also report a within-cell correlation ρ_w: average ranks are computed within each fixed cell, centered within that cell, and combined across episodes using Pearson correlation, so between-cell differences do not contribute; the range of the individual cell correlations is shown beside it. For the hidden-state readout, pooled and within-cell estimates are negative or near zero in every arm. For the logit readout, the pooled correlation is positive in each synthetic arm, but the pattern is not uniform within cells: for Mistral, the pooled value is 0.200 whereas the four cell correlations are -0.399, 0.041, -0.006, and -0.215. IOI has small pooled estimates for both readouts, which do not establish absence of association. All correlations are retrospective and descriptive; the eight pooled p-values are unadjusted, and no tests are reported for the within-cell estimates.

Table 7: Pooled Spearman and aggregate within-cell rank correlations between episode-level projections and donor-oriented margin movement (U_i^m for synthetic arms, cross-only; $U_i^{m,\mathrm{net}}$ for IOI), retrospective; $n = 512$, 256, 256, and 500. *Pooled*: across the fixed task cells, with two-sided unadjusted p-values in parentheses. *Within-cell* ρ_w: Pearson correlation of average ranks computed and centered within each cell (four cells for synthetic arms, ten for IOI); brackets give the range of the individual cell correlations. IOI logit values use centered logits.

	Hidden-state projection		Logit-contrast projection	
Arm	Pooled ρ (p)	Within-cell ρ_w [cells]	Pooled ρ (p)	Within-cell ρ_w [cells]
Qwen SAME/DIFFERENT	−.147 (.0009)	−.112 [−.384, .086]	.743 ($< 10^{-80}$)	.235 [−.052, .630]
Mistral SAME/DIFFERENT	−.235 (.0001)	−.359 [−.446, −.293]	.200 (.001)	−.145 [−.399, .041]
Qwen DAX/WUG	−.163 (.009)	−.118 [−.387, .013]	.344 ($< 10^{-7}$)	−.052 [−.524, .411]
Qwen IOI	−.068 (.13)	.035 [−.214, .213]	.051 (.26)	.095 [−.226, .541]

Leave one template out. Table 8 recomputes the three aggregate IOI endpoints after excluding each evaluation template in turn. Every estimate keeps its sign. No single included template therefore drives the aggregate direction; the check does not establish generalization to unseen template families.

Table 8: IOI endpoints with one evaluation template excluded (retrospective; 400 episodes each).

Excluded template	Hidden $\widehat{\Delta}^h$	Centered logit $\widehat{\Delta}^{\tilde{\ell}}$	Margin $\widehat{\Pi}^m_{\mathrm{net}}$
10	.96612	.00551	−.01541
11	.96881	.00838	−.03584
12	.97035	.00836	−.05376
13	.96839	.00511	−.03337
14	.96774	.00885	−.02516
None (all 500)	.96828	.00724	−.03223

Directional margins and candidate probabilities. Table 2 reports each IOI patch direction separately. Two-candidate probabilities are $\sigma(m)$ and match the archived probabilities to within 10^{-7}; the largest single change is 0.84 percentage points. These are two-candidate quantities and do not measure full-vocabulary probability change.

Algorithm 3 Readout-wise donor projections with stratified bootstrap (retrospective analysis of saved records)

Input: records with $\Delta_i^h, \Delta_i^\ell$, margins $m_{i,E}, m_{i,D}$, movement U_i^m (or patched and same-format margins), cells c_i
Output: readout-wise estimates, 95% intervals, scale summaries, and robustness checks

$\widehat{\Delta}^v \leftarrow n^{-1} \sum_i \Delta_i^v$ for $v \in \{h, \ell\}$ ▷ *archived intervals kept*
for $i = 1, \ldots, n$ **do**
  $g_i \leftarrow m_{i,D} - m_{i,E}$
  if same-format margins are archived **then**
    $U_i \leftarrow U_i^{m,\text{net}}$ ▷ *Eq.* (20)
  else
    $U_i \leftarrow U_i^m$ ▷ *cross-only*
$\widehat{\Pi}^m \leftarrow \sum_i U_i \,/\, \sum_i |g_i|$; record $\overline{|g|}$
rng $\leftarrow$ `default_rng`(20260925)
for $b = 1, \ldots, 20{,}000$ **do**
  $I_b \leftarrow [\,]$ ▷ *a list: repeated indices are kept*
  for each cell c in sorted order **do**
    $I_b \leftarrow \text{concat}\left(I_b,\ \text{SAMPLEWITHREPLACEMENT}(\{i : c_i = c\},\ n_c)\right)$
  $\Pi_b \leftarrow \sum_k U_{I_b[k]} \,/\, \sum_k |g_{I_b[k]}|$ ▷ *sum over list positions*
interval $\leftarrow$ 2.5th and 97.5th percentiles of $\{\Pi_b\}_b$
robustness: exclude each IOI template in turn; per patch direction, margin changes, sign flips, $|\Delta\sigma(m)|$
return $\widehat{\Delta}^h, \widehat{\Delta}^\ell, \widehat{\Pi}^m$, interval, median and mean of $R_i = U_i/|g_i|$ over $g_i \neq 0$, $\#\{U_i > 0\}$

Table 9: Post hoc counts with positive correct-answer margin, by cell. $c = 0$ means SAME or DAX↦X; $c = 1$ means DIFFERENT or DAX↦Y. For SD, x is the input bit; for DW, $x = 0$ is DAX and $x = 1$ is WUG.

Arm	c	x	Examples	Direct	Direct-format ties
Qwen SD	0	0	0/128	0/128	0
	0	1	0/128	0/128	0
	1	0	128/128	128/128	0
	1	1	128/128	128/128	0
Mistral SD	0	0	64/64	64/64	0
	0	1	64/64	64/64	0
	1	0	0/64	0/64	0
	1	1	27/64	34/64	1
Qwen DW	0	0	63/64	56/64	3
	0	1	7/64	34/64	0
	1	0	63/64	49/64	0
	1	1	1/64	19/64	0

E Additional descriptive results from saved records

E.1 Candidate preference by task cell

Table 9 decomposes the aggregate preference rates in Table 1 and supports the candidate-preference description in Section 4.3. A cell contains repeated nonce draws with fixed mapping/query content; it is not a sample of different semantic tasks. Exact ties occur only in the direct-format cells marked below. No confidence intervals over a broader task population are claimed.

Mean restricted correct-candidate probabilities are 0.53559/0.52864 for primary Qwen, 0.55174/0.55080 for Mistral, and 0.52872/0.55799 for DAX/WUG, in examples/direct order. A mean probability above 0.5 can coexist with only 50% strict preference because confidence differs across correct and incorrect cells. These numbers should not be interpreted as calibrated probabilities of generated task success.

Table 10: Descriptive OLMo decomposition of the recorded net logit endpoint. Sample SD is across nonce-varied episodes; it is not uncertainty across task families.

Rule	Bit	n	Mean Δ^{ℓ}	Sample SD	Positive episodes
SAME	0	64	.03352	.03007	55/64
SAME	1	64	−.35703	.05377	0/64
DIFFERENT	0	64	−.20517	.35047	17/64
DIFFERENT	1	64	.66015	.80991	47/64

Table 11: Paired analysis of unpatched candidate preference. A format is positive when its correct-answer margin exceeds zero. Nonpositive margins include ties; the final column counts episodes with a tie in either format. Rows sum to n across the four joint outcomes.

Arm	Both positive	Examples only	Direct only	Neither	Any tie
Qwen SD	256	0	0	256	0
Mistral SD	155	0	7	94	1
Qwen DW	113	21	45	77	3

E.2 OLMo condition heterogeneity

OLMo uses arrow-formatted examples and `Direct rule:` histories rather than frozen template 3. It samples the control nonce independently without enforcing inequality. These differences are part of the separate extension apparatus. OLMo's hidden cross-format mean is 0.58816, with same-format mean 0.00095 and net mean 0.58722. Its logit cross-format mean is 0.03700, same-format mean 0.00414, and net mean 0.03286. All recorded hidden/logit bridge-cut distances are zero. The extension smoke log contains endpoint observations overlapping the first four final episodes. It is therefore presented as a later extension rather than an independent untouched confirmation of every observation. Fresh held-out seeds are required for a new confirmatory follow-up.

E.3 Candidate-position patterns and paired baselines

In primary Qwen SAME/DIFFERENT, the preferred candidate is $1 - x$ in all 512 episodes under both formats. This equals the answer in the second mapping definition, DIFFERENT means answer $\{1\text{-x}\}$. It is not necessarily the last digit anywhere in the query, which ends with Reply only 0 or 1. Mistral prefers x in all 64 episodes of each of the first three cells. In DIFFERENT/1, it prefers x in 37/64 examples-format episodes and 29/63 strict direct-format preferences, with one direct-format tie. For DAX/WUG, the examples baseline prefers the first-listed target a in 246/256 episodes and the direct baseline in 180/256. These descriptive alignments resemble documented recency and option-order biases (Zhao et al., 2021; Pezeshkpour & Hruschka, 2024), but the favored position differs across arms (second clause for Qwen SAME/DIFFERENT, first-listed target for DAX/WUG, input bit for Mistral). An equally consistent reading is a rule-independent default within each arm, for example answering as if the rule were always DIFFERENT. In SAME/DIFFERENT these readings are confounded because DIFFERENT is always stated second. A counterbalanced statement-order test separates three outcomes: preferences that flip with order (position), preferences that stay on the same wrong branch (fixed default), and preferences that track the stated rule (competence).

E.4 Component means and secondary outcomes

The secondary signed margin and two-candidate probability measures average the two patch directions and do not subtract same-format movement. Their equivalence criteria apply to these signed means, not to per-episode accuracy or absolute changes.

Table 12: Descriptive control means and unpatched format distances from the saved records. C, W are the bidirectionally averaged cross-/same-format projections. d_h, d_ℓ are mean Euclidean baseline distances; their scales are model dependent.

Arm	C^h	W^h	C^ℓ	W^ℓ	d_h	d_ℓ
Qwen SD	.68846	.00099	.25874	.00041	1.25234	272.95789
Mistral SD	.80283	.00184	.05953	.00088	.03578	13.79986
Qwen DW	.60620	.00105	.04723	.00005	1.05629	218.49178
OLMo SD	.58816	.00095	.03700	.00414	.48391	152.15451

Table 13: Frozen secondary donor-oriented means and 90% intervals. Equivalence refers only to the specified average endpoints. Unlike Eq. (7), these endpoints do not subtract same-format movement (Section C).

Arm	Margin mean [90% CI]	Probability mean [90% CI]	Both pass?
Qwen SD	.45166 [.43629, .46694]	.00562 [.00467, .00658]	No
Mistral SD	.01162 [.01093, .01231]	.00232 [.00216, .00248]	Yes
Qwen DW	$-.04874$ [$-.05872$, $-.03870$]	$-.01116$ [$-.01320$, $-.00912$]	Yes

E.5 Distributions and finite-sample decomposition

For arm a, the empirical cumulative distribution is

$$\widehat{F}_a(t) = n_a^{-1} \sum_{i=1}^{n_a} \mathbf{1}[\Delta_{a,i}^{\ell} \le t]. \tag{15}$$

Empirical quantiles describe observation spread, not uncertainty about the mean. The Qwen primary, Mistral, DAX/WUG, and OLMo raw-logit medians are respectively 0.25880, 0.05792, 0.04988, and -0.02087. Their 5th–95th percentile ranges are [0.11686, 0.37957], [0.03242, 0.08616], [-0.01561, 0.09994], and [-0.55741, 1.37605].

Let g index the four OLMo task cells, $w_g = n_g/n$, and $\bar{\Delta}_g$ their means. The exact finite-sample variance decomposition is

$$\frac{1}{n}\sum_i (\Delta_i - \bar{\Delta})^2 = \sum_g w_g \frac{1}{n_g}\sum_{i\in g}(\Delta_i - \bar{\Delta}_g)^2 + \sum_g w_g(\bar{\Delta}_g - \bar{\Delta})^2. \tag{16}$$

The second term accounts for 43.8732% of the total in the OLMo records. This identity describes fixed-cell heterogeneity; it does not identify a causal interaction.

F Scope of the contrast and proofs of the formal qualifications

What the design identifies. Eq. (3) is a causal effect of replacing retained bridge states in a constructed cache, with parameters, query, positions, and mask held fixed. Three limits follow from the construction. First, it is not the total effect of the original prompt: the prefix is deliberately replaced and donor and recipient blocks are combined in a way that need not occur in an ordinary forward pass. Second, it is not a natural mediation decomposition of an unmodified conversation or a deletion guarantee for a deployed system. Third, it does not identify which property of the source text causes the effect. The two formats convey the same intended task but differ in wording, length, repetition, and syntax, and a nuisance control removes only its specified alternative. Subtracting the same-format coefficient asks whether the cross-format replacement moves the output further along the baseline axis than the implemented control; it makes the coefficient interpretable within an episode but does not equate interventions across models.

The propositions below are mathematical observations about the measurement design. They are not additional model experiments.

Proposition 1 (Projection is incomplete as a distance measure)**.** *Let $d = b - a \neq 0$. Any displacement can be written $z - a = \alpha d + u$ with $u^\top d = 0$. Then $P(a, z, b) = \alpha$, while $\|z - b\|_2^2 = (\alpha - 1)^2\|d\|_2^2 + \|u\|_2^2$. Therefore $P = 1$ does not imply $z = b$, and $P = 0$ does not imply no intervention effect.*

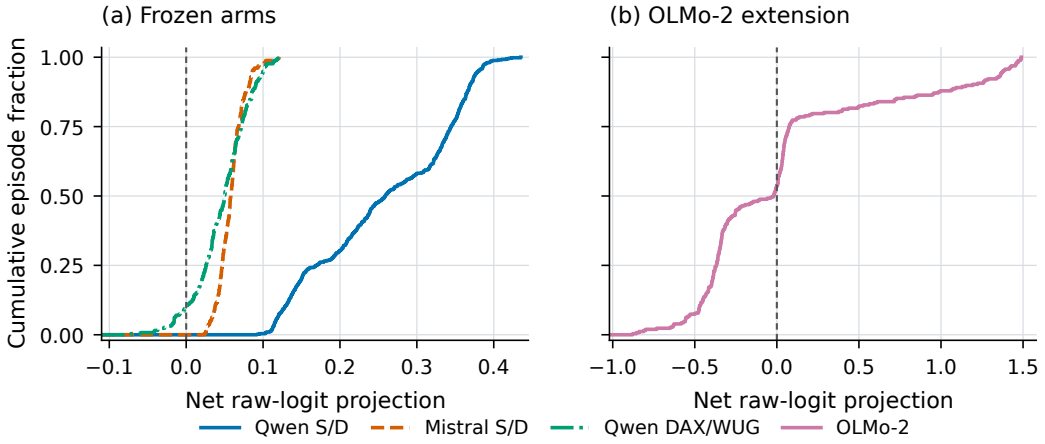

Figure 6: Empirical cumulative distributions of all recorded bidirectional net-logit projections. Each episode contributes one value; the ordinate gives the fraction at or below the abscissa. Frozen arms and the OLMo extension are shown separately for readability. The zero reference distinguishes positive and nonpositive movements. Curves show observation spread, not confidence bands.

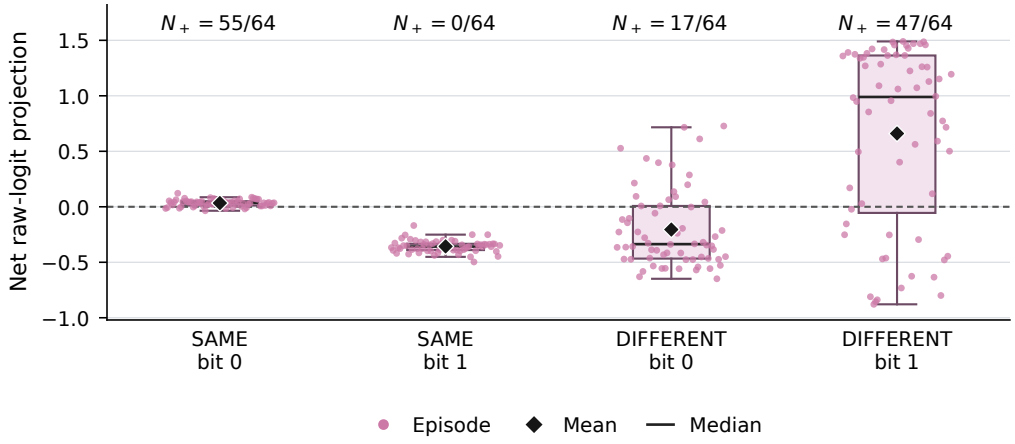

Figure 7: Actual OLMo episode-level net-logit measurements, with 64 records per condition. Horizontal jitter separates overlapping points. Boxes span the interquartile range (IQR); center lines show medians, diamonds show means, and whiskers extend to observations within 1.5 IQR of the box. These are distribution summaries, not inferential intervals.

Proposition 2 (Raw-logit movement need not imply probability change)**.** *For logits $\ell \in \mathbb{R}^V$ and any scalar k,* $\mathrm{softmax}(\ell + k\mathbf{1}) = \mathrm{softmax}(\ell)$. *Nevertheless, the raw-logit projection can be nonzero when recipient, donor, and patched outputs differ only by such offsets.*

Proposition 3 (Equality after complete bridge masking)**.** *Suppose two deterministic evaluations have identical model parameters, current-query tokens and positions, accessible source-prefix K/V, and all other accessible state. Suppose every differing cached entry lies at bridge positions, and every bridge position is masked from the current query at every layer. Then the two query computations and their final outputs are identical.*

$$R_i^v = \frac{\|Y^v_{E,\mathrm{cut}} - Y^v_{D,\mathrm{cut}}\|_2}{\|Y^v_E - Y^v_D\|_2}, \qquad v \in \{h, \ell, \widetilde{\ell}\}, \tag{17}$$

where the centered-logit version is recorded for IOI. Denominators must be nonzero.

Proof of Proposition 3. The initial current-token embeddings and positional inputs are identical. At the first layer, each query receives identical accessible keys and values: source entries match, bridge entries are inaccessible, and the current-query entries derive from identical inputs. Attention, residual updates, and feed-forward operations therefore return identical outputs. Induction gives equality at each subsequent layer and at the output head. For a multi-token current query, the same argument applies with the same causal mask and identical preceding current-token states. Thus the numerator of Eq. (17) is zero. The claim presumes deterministic execution and complete masking; it does not assert that arbitrary implementations have these properties. □

Proof of Proposition 1. Substitution of $z - a = \alpha d + u$ into Eq. (4) removes the orthogonal term, giving $P = \alpha$. Expanding $z - b = (\alpha - 1)d + u$ gives $\|z - b\|_2^2 = (\alpha - 1)^2\|d\|_2^2 + \|u\|_2^2$ because the cross term is zero. □

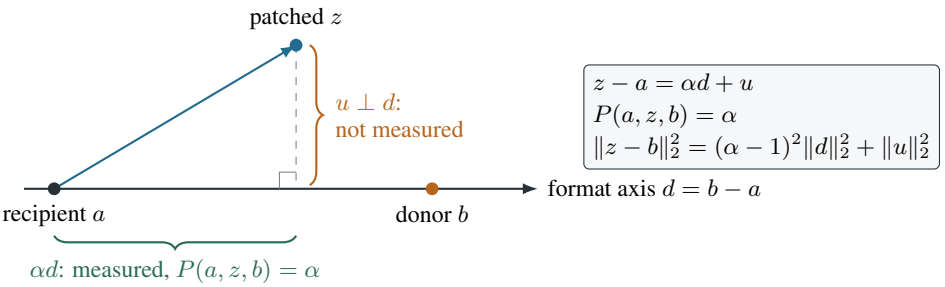

Figure 8: Geometry of the donor projection (Proposition 1). The patched output z is decomposed into a component αd along the recipient–donor axis and an orthogonal component u. The projection reports only α, so $P = 1$ does not imply $z = b$ (the distance to the donor still contains $\|u\|_2^2$, as the boxed identity shows), and α can lie outside $[0, 1]$. Measuring both components requires the full output vectors.

Proof of Proposition 2. The common factor e^k cancels in each normalized probability. For an explicit counterexample, take $a = (0, 0)$, $b = (1, 1)$, and $z = (0.7, 0.7)$. Equation 4 equals 0.7, while all three softmax distributions are $(1/2, 1/2)$. This demonstrates a limitation of the metric; it does not show that the observed experimental effects consist only of common offsets. □

Markers satisfy the estimand. For a partial mixture, define

$$z_\lambda = (1 - \lambda)a + \lambda b, \quad 0 < \lambda \leq 1. \tag{18}$$

Substituting $z_\lambda - a = \lambda(b - a)$ into Eq. (4) gives $P(a, z_\lambda, b) = \lambda$ regardless of why a and b differ. Under this mixture construction, a format marker alone can yield a positive net coefficient even when it contains no task-relevant information: if the patched output mixes the two baselines and a same-format donor carries the recipient's marker, the cross-format coefficient is positive while the same-format coefficient is near zero. The construction is a possibility result, not a claim that every marker produces the effect. A positive projection is therefore compatible with, but not evidence for, a task-relevant trace; surface-matched donors are needed to distinguish the two.

G Artifact provenance and executable reanalysis

G.1 Separate evidence sources

The results comprise three frozen synthetic arms, a later OLMo extension, a later IOI extension, and retrospective analyses of their records. Candidate-accuracy diagnostics of the original arms, candidate-position descriptions, paired baseline tables, distribution plots, and IOI condition plots are retrospective. They are not additional preregistered confirmations. Answer-margin projections, directional margin summaries, and robustness checks (Section D) are also retrospective. The IOI

protocol contains decision criteria, but its prospective lock is not independently authenticated by the archived provenance (Section H.3).

Original frozen records contain bidirectional geometric summaries, cut distances, baseline answer scores, and selected secondary endpoints. They do not retain full vectors or all direction-specific patched scores. IOI adds centered-logit summaries and separate patched candidate margins, but does not retain the full vectors, direction-specific geometric projections, or cut-condition answer margins. Saved-data analysis therefore verifies reported aggregate arithmetic; recreating the projections from tensors requires model inference.

G.2 Included files and commands

The supplementary archive contains the decoded frozen and OLMo records under `evidence/`, the IOI records, runner, and analysis under `ioi/`, and scripts for CPU reanalysis and empirical figure generation. The IOI visualization script reports its exploratory cell-bootstrap seed separately from the original primary-analysis seed. The primary endpoint intervals are preserved.

```
latexmk -pdf -interaction=nonstopmode -halt-on-error new.tex
```

CPU statistics can be recomputed from the saved records without model weights. The IOI GPU runner emits JSON records to standard output rather than writing the analysis input file directly; a complete rerun workflow needs a record exporter, the pinned model, and compatible GPU dependencies. The original source package also lacks an independently timestamped protocol lock. These are disclosed reproducibility limitations, not replaced with inferred provenance.

G.3 Preserved development outcomes

The earlier project record reports a biased verbal source-attribution instrument, steering without selective answer-logit effects beyond a matched random perturbation, an untrained bridge that did not reliably control answers, and early task-performance failures. A Phi attempt failed at the cache interface; tokenizer failures invalidated other launches. These outcomes informed the design and are not pooled with completed experimental arms. OLMo's smoke log overlaps four final episodes, while the frozen evaluation template appeared in development. The IOI development/final split is described separately in Section H.1.

H IOI validation: construction, analysis, and provenance

This extension evaluates presentation-history effects during indirect-object-identification (IOI) candidate discrimination. It preserves source neutralization, the fixed bridge, and early-layer bridge interchange, while introducing persistent padding masks and a mean-centered logit endpoint. These amendments define a separate experiment; they do not retrospectively repair the frozen arms or isolate the cause of their lower candidate performance.

H.1 Episode construction and recorded protocol

The runner defines 15 sentence templates. Templates 0–9 supply source examples; templates 10–14 supply evaluation queries. In each query, A is the intended indirect object and B is the repeated subject. The two orders are BABA and ABBA, including the intended final completion: changing order exchanges only the first occurrence of "B and A". The five evaluation stems are:

10: After lunch, B and A went to the PLACE. B gave a OBJECT to
11: Afterwards, B and A went to the PLACE. B gave a OBJECT to
12: Then, B and A had a long argument. Afterwards B said to
13: The PLACE B and A went to had a OBJECT. B gave it to
14: Friends B and A found a OBJECT at the PLACE. B gave it to

These are literal construction patterns before substitution, including the article "a". The completion query contains no restated IOI rule. Places are store, garden, office, school, station, market, theatre,

and museum; objects are book, drink, letter, gift, snack, ticket, ball, and note. Template 12 does not use the sampled place or object.

Names are eligible only when their space-prefixed form is one token under the pinned tokenizer; the runner additionally checks the complete query-plus-candidate boundary. The 24 source names are Alice, Bob, Carol, David, Emma, Frank, Grace, Henry, Irene, Jack, Karen, Louis, Megan, Nathan, Olivia, Peter, Rachel, Simon, Tina, Victor, Wendy, Arthur, Bella, and Colin. The 28 eligible final names are Laura, Martin, Nina, Oscar, Paula, Robert, Sarah, Thomas, Vincent, William, Xavier, Brenda, Dennis, Elaine, George, Hannah, Jason, Linda, Marcus, Nicole, Patrick, Rebecca, Samuel, Teresa, Walter, Benjamin, Catherine, and Daniel. The separate pilot pool is Diana, Edgar, Fiona, Gavin, Helen, Isaac, Julia, and Kevin, subject to the same filter. Evaluation names are disjoint from source and pilot names. Evaluation templates are excluded from demonstrations but also used by the pilot; "held out" therefore refers to source demonstrations, not to investigator-unseen templates.

For episode i, source variant $v = i \bmod 4$ selects four examples. Example $j \in \{0, 1, 2, 3\}$ uses template $(2v + j) \bmod 10$, place index $(v + j) \bmod 8$, object index $(2v + j) \bmod 8$, and places the indirect object first when $(v + j) \bmod 2 = 1$. Eight distinct source names are sampled for each examples history, giving four distinct name pairs. Its header is `Session s.  Examples of sentence completion:`; each completed example ends in the correct name and a period. A direct history has header `Session s.  Direct task rule:` and selects one of four fixed paraphrases by v. These specify the recipient or non-repeated name, distinguishing it from the repeated giver or subject; exact strings are retained in the runner.

Session identifiers are four contiguous decimal digits, including leading zeros. The same-format donor uses a distinct identifier. The examples donor independently resamples its eight names while retaining the same variant; the direct donor changes the identifier but retains the rule wording. Thus the examples control changes both names and identifier. Cross-format donors share the recipient identifier. The neutral source is `Session s.  Earlier material was neutral: cloud stone window copper river lamp paper chair.`, followed by a newline.

The final seed is 9028173. Fifty cycles independently shuffle the ten template/order cells, yielding 50 episodes per cell. Query names are sampled without replacement within an episode; place and object are sampled uniformly. The resulting 500 episodes contain 497 distinct query strings, with three repeated queries under different histories. The pilot configuration specifies 40 episodes and seed 184731; its observations are not pooled with final records.

The protocol declares thresholds of 80% candidate accuracy for query-only, each neutralized format, and pooled cross-format patches, and 75% for each name order. Both primary projection intervals must exclude zero, cross-format means must exceed same-format means, and bridge-cut distances must be at most 10^{-6} of uncut distances. The archived records satisfy every criterion, and we report the underlying counts rather than only the pass flags. The criteria answer a bounded question, whether a retained-history effect is present when the model discriminates the answer reliably. They do not require that the histories teach IOI, that the bridge is necessary for answering, or that the donor contrast is surface-matched. The 40-episode pilot used a disjoint name pool and the same evaluation templates.

Algorithm 4 IOI episode construction and measurement as implemented

Input: pinned model and tokenizer; RNG state; episode index i; evaluation template τ; name order o
Output: one scalar record with margins, contrasts, distances, and query metadata

sample query names, place, and object; identifiers $s \neq s'$; $v \leftarrow i \bmod 4$
build $E(s), D(s), E(s'), D(s')$ with variant v, and *neutral*(s), *neutral*(s')
for each history x **do**
    reject if $|\mathrm{tok}(x)| > 192$; pad to 192 with EOS; append the 5-token bridge
    prefill at explicit positions $0, \ldots, 196$ with filler masked
neutralize by Eq. (1), keeping the neutral padding mask ▷ $E(s), D(s)$ *share* $N(s)$
evaluate query-only (ordinary positions), intact E, D, and neutralized E, D (positions from 197)
for each $(r, d) \in \{(E, D), (D, E), (E, E'), (D, D')\}$ **do**
    evaluate $I_{\mathcal{L}}(C_r, C_d)$ with $\mathcal{L} = \{0, \ldots, 5\}$ on bridge positions 192–196
evaluate neutralized E, D with every bridge position masked at every layer ▷ *cut*
read hidden entry 8, raw and centered logits (Eq. (19)), and the one-token IO-minus-subject margin
return projections by Eqs. (4) and (7); distances; margins (not for cut); probabilities; metadata

Table 14: Descriptive centered-logit outcomes by template and name order. Each cell contains 50 episodes; counts use strict positive net movement.

	ABBA: correct name first		BABA: correct name second	
Template	Mean	Positive	Mean	Positive
10	.013631	32/50	.014690	35/50
11	.003038	29/50	.002332	28/50
12	.006315	33/50	−.000722	26/50
13	.014278	35/50	.017266	36/50
14	−.003718	24/50	.005326	26/50

H.2 Statistical analysis and condition-level outcomes

Candidate success is $\mathbf{1}[m_i > 0]$, where m_i is the correct-name logit minus the repeated-subject logit. This is not a full-vocabulary argmax or generated-answer measure. Query-only succeeds in 499/500 episodes; intact and neutralized formats and each patch direction succeed in 500/500. Each of the latter conditions succeeds in 250/250 episodes per name order. The two patch directions share episodes: report them separately or score joint success per episode, rather than treating 1,000 outcomes as independent trials. The conventional Wilson interval for 500/500 episode successes is [99.24%, 100%], conditional on this evaluation construction.

The archived analysis uses seed 771903 and 100,000 stratified bootstrap resamples, retaining 50 observations in each of ten template/order cells. Percentile intervals estimate variation within this fixed template set, not uncertainty across arbitrary templates, models, or task families. Centered logits are

$$\widetilde{\ell} = \ell - \frac{\mathbf{1}^\top \ell}{V}\mathbf{1}, \tag{19}$$

which remove common offsets that leave softmax probabilities unchanged. Hidden cross/same means are 0.972265/0.003984; centered-logit means are 0.007476/0.000232. Their net means and 95% intervals are 0.968281 [0.966883, 0.969670] and 0.007244 [0.004898, 0.009612]. Raw-logit net movement, retained for comparison, is 0.004404 [0.001879, 0.006913]. The analysis also computes 100,000 sign-flip draws; these require sign-exchangeability under their reference null and are not design-based randomization tests.

Hidden net movement is positive in all episodes; centered-logit net movement is positive in 304/500. Table 14 and Figure 9 expose the output heterogeneity, including two negative cell means. All recorded hidden and centered-logit cut distances are zero. Cut-condition answer margins were computed transiently but not archived, so their accuracy cannot be recovered from this packet.

The answer-margin analysis asks whether the task-answer margin itself moves toward the opposite-format baseline. With $s_i = \mathrm{sgn}(m_{i,D} - m_{i,E})$, define

$$U_i^{m,\mathrm{net}} = \frac{s_i}{2}\left[(m_{i,E\leftarrow D} - m_{i,E\leftarrow E'}) - (m_{i,D\leftarrow E} - m_{i,D\leftarrow D'})\right]. \tag{20}$$

The saved records give mean $U_i^{m,\mathrm{net}} = -0.0178$ logits against a mean format gap of 0.552 logits, a gap-weighted projection of -0.032 $[-0.044, -0.020]$ (Section D). This post hoc endpoint was not part of the stated primary decision rule and does not invalidate the primary endpoints. It shows that a hidden-state donor-axis projection near one and a positive centered-logit projection need not imply donor-directed movement of the answer margin. No patched candidate choice changes in either direction (Table 2).

H.3 Runtime, archived fields, and provenance limits

The run uses `Qwen/Qwen2.5-3B-Instruct`, revision `aa8e72537993ba99e69dfaafa59ed015b17504d1`, with float16 weights, SDPA attention, evaluation mode, and inference mode. Recorded software is Python 3.11.2, PyTorch 2.6.0+cu124, CUDA 12.4, and Transformers 5.13.0 on an NVIDIA A10G. The bridge is `History processed. Ready.\n`, token IDs [13424, 15233, 13, 30982, 624]. The job record reports completion on 19 September 2026.

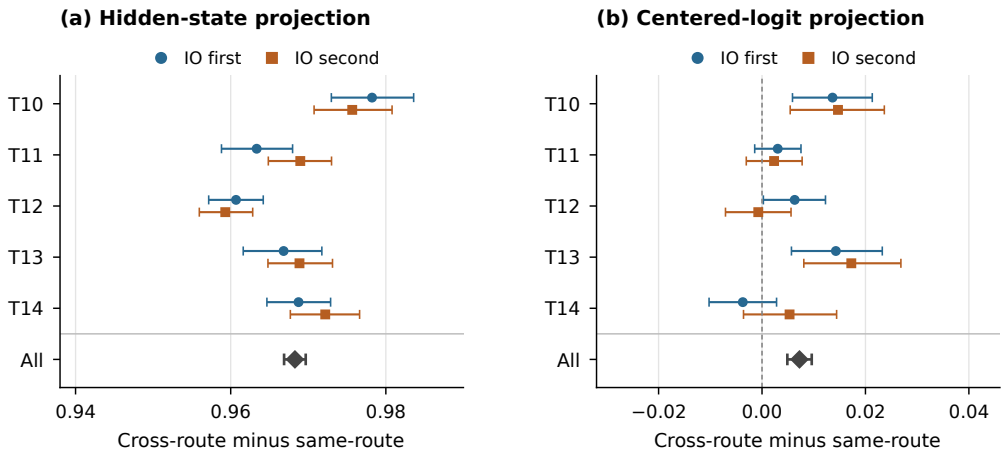

Figure 9: IOI effects by template and name order, calculated from the 500 saved records. Panel (b) repeats Figure 4b to allow side-by-side comparison with the hidden-state panel; it is not additional evidence. Points are cell means; bars are exploratory within-cell percentile bootstrap intervals. The ten cells are fixed evaluation settings, and these intervals are not multiplicity-adjusted tests of template differences.

The protocol states that it was fixed before the final run. The archived provenance record, however, replaces the protocol and runner hashes with `<redacted>`, and the pilot outputs and final job log are not archived. The files therefore support reproducibility of the saved-record analysis but do not independently establish when the protocol was locked. We accordingly describe this arm as an extension with stated prospective criteria, not as externally verified preregistration.

The packet contains 500 scalar records, runner and analysis code, protocol, figures, and checksums. Reanalysis reproduces the archived aggregates without model inference. Regenerating activations requires the pinned model and compatible GPU environment. The runner emits JSON records and a compressed record bundle to standard output; reproducing the analysis requires retaining and decoding that output into `results_records.json`. File checksums establish content consistency, not when results were inspected. Full activation vectors, individual directional projections, full probability distributions, generated answers, and cut-condition answer scores are absent; they must not be inferred from the stored aggregate projections.

I Proposed tests of specificity and persistence

The IOI extension implements candidate-level competence checks, balanced name order, persistent padding masks, and centered-logit endpoints. The remaining comparisons below are proposed extensions. They have not been performed, and no successful outcome is asserted for them. Their purpose is to distinguish explanations that the present evidence leaves open.

Competence gates come first. Establish, on development data, that the apparatus can use the rule: chat templates applied, candidate answers not stated inline or with their order counterbalanced, and both token-candidate and generated-answer scoring in untouched, query-only, neutralized-history, and bridge-cut conditions. Prespecify a competence threshold before any patching. For the candidate-position pattern of Section 4.3, the discriminating manipulation is reordering the candidate statements, which separates three outcomes: preferences that flip with order (position), preferences that stay on the same wrong branch (fixed default), and preferences that track the stated rule (competence). Scoring candidates both with and without a separating space after `Answer:` tests whether the scoring position contributes to the failure.

Check structural invariants. A self-swap using identical donor and recipient caches should reconstruct the original cache exactly; a zero-layer patch should leave outputs unchanged; a full-layer bridge swap with a matched neutral prefix should reproduce the opposite baseline under identical query inputs; complete masking should yield the Proposition 3 zero. These are invariants of a compatible implementation, not discoveries about the model, and they complement tokenization and competence checks. To isolate the padding-mask contribution in the original synthetic tasks, a dedicated arm would compare original later mask versus a persistently preserved neutral-prefix padding mask.

Donor controls are surface-matched and factorial. Cross presentation format (examples versus instruction) with task content (same versus different task), matching length and repetition as closely as possible, so that format and content effects and their interaction can be estimated. Cross-format donors should also be compared against paraphrase donors matched in length and repetition, irrelevant-format histories, and corrupted demonstrations, to test whether the effect extends beyond the matched lexical and structural differences. A claim that retained states transfer task information additionally requires a source-dependent task: place an arbitrary mapping only in the source, verify high full-history and low query-only accuracy, and then test the retained bridge states. Crossing format and nonce changes (cross format/same nonce, cross format/new nonce, same format/same nonce, same format/new nonce) separates format and nonce effects and their interaction, which the single implemented control cannot. An untrained or otherwise task-incompetent model is a useful *specificity* arm: a geometric effect there would show the metric does not require competent rule use; its absence would not certify every trained-model effect.

Outcomes span geometry, probability, and behavior. Retain full output vectors, both interchange directions, and patched candidate choices. Alongside Eq. (4), report centered logits and a probability distance,

$$\mathrm{TV}(p, q) = \tfrac{1}{2} \sum_{v=1}^{V} |p_v - q_v|, \tag{21}$$

together with candidate margins and generated decisions, so that geometric change, probability change, and task change are distinguished rather than conflated. Centering removes common offsets; total variation measures probability change but by itself does not establish task relevance or donor specificity.

Control patch location and readout depth. This is the most consequential remaining test. Keep the model, queries, source construction, masks, and donor construction of the IOI arm fixed, and compare equal-width early, middle, and late patch blocks. From each forward pass, read hidden states at every layer, record the two interchange directions separately, and measure centered logits, a full-distribution distance such as total variation (Eq. (21)) or KL divergence, answer margins, and candidate choices. Include identity, zero-layer, and all-layer swaps as implementation checks; with a shared neutral prefix and an identical query, the all-layer swap must reproduce the donor baseline exactly, an invariant rather than a finding. If large projections follow proximity to patched layers, the result would support a substantial contribution from intervention location. If the dissociation persists across the prespecified placements and depths, it would demonstrate robustness to those choices while leaving downstream sensitivity, hybrid-cache effects, and task specificity unresolved. The existing arms offer no test: hidden-state projections do not follow the fraction of donor-reading layers before the readout (0.80 at five of seven for Mistral; 0.59–0.97 at six of eight elsewhere), but model and task differ between arms.

Measure persistence and interference. Prespecified delay lengths, distractor text, competing rules, and fresh-state controls estimate decay and interference. Uncertainty should be estimated at the level of independently varied task and template families, with nonce variation nested inside families, and both aggregate and condition-level outcomes reported.

Table 15 maps each open ambiguity to the comparison that would resolve it; Algorithm 5 fixes the confirmation sequence, including disjoint development/confirmation splits and thresholds frozen before results are observed. Fresh-seed confirmation must not reuse endpoint-observed smoke episodes (Section E).

Table 15: Targeted follow-ups and the ambiguity each would resolve.

Question	Required comparison	Discriminating observation
Can the apparatus use the rule?	Query-only, intact history, neutralized history, and cut; score token candidates and generated answers	Locate where competence is lost before interpreting format effects
Is candidate position the failure mode?	Counterbalanced order of candidate statements; chat template on versus off; candidates scored with and without a separating space	Flip with order (position); stay on one wrong branch (fixed default); track the stated rule (competence)
Is the hidden-state magnitude structural?	Early, middle, and late patch blocks of equal width; all-layer hidden readouts; both directions; centered logits, full-distribution distance, margins, choices; identity, zero-layer, and all-layer invariants	Projections that follow proximity to patched layers support a location effect; dissociation at every placement shows robustness to placement but leaves downstream sensitivity and task specificity open
Does the effect carry task information?	Arbitrary mapping present only in the source; high full-history and low query-only accuracy; then bridge interchange	Donor-directed answer change only if retained states carry the source's task content
Is padding responsible?	Original later mask versus persistent neutral-prefix padding mask	Attribute a difference only if it changes under this controlled replacement
Is format specificity semantic?	Paraphrases, matched lengths/repetition, irrelevant-format histories, and corrupted demonstrations	Effects generalize beyond lexical markers while tracking the intended relation
Does donor matching matter?	Cross format/same nonce, cross format/new nonce, same format/same nonce, same format/new nonce	Estimate separate format and nonce effects and their interaction
Is output behavior affected?	Raw/centered logits, full probabilities, candidate margins, generated choices	Distinguish geometric changes from probability and task changes
Does history persist?	Prespecified delay lengths, distractors, competing rules, and fresh-state controls	Estimate decay and interference across genuinely new task families

Algorithm 5 Proposed validation sequence for a new confirmation (not executed)

Input: task and template families; model and software revisions
Output: confirmatory estimates with decisions fixed before results are seen

Development phase

1: define families and disjoint development and confirmation splits
2: choose task-competence criteria on development data only
3: verify boundaries, cache shapes, source equality, identity-swap and all-layer invariants, and cut isolation
4: freeze prompt format, masks, layers, endpoints, controls, exclusions, allocation, bootstrap clusters, seeds, revisions

Confirmation phase

5: **for each** eligible confirmation cell **do**
6: run every condition; keep full or sufficient summaries, both directions, choices, and perturbation norms
7: estimate uncertainty over families, with nonces nested inside families
8: report aggregate and condition-level outcomes, clean negative findings, and setup failures separately