TEXTUAL SALIENCE TIPS THE BALANCE: MODULATING LONG-CONTEXT MEMORY RETRIEVAL IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable retrieval from long contexts is fundamental for LLM agents to reuse interaction histories and prior decisions. Yet even when target information remains available, retrieval can fail when similar memories compete. We ask whether short textual salience cues can shift retrieval priority among competing memories. To isolate this effect from world knowledge and semantic associations, we construct a controlled key–value framework with random bindings and similar-key distractors, evaluating five open-weight LLMs. We test three affectively framed cues—threat, reward, empathy—together with a non-valenced importance cue. Most errors match distractor values already present in context, indicating confusion among competing bindings rather than simple forgetting. Under this interference, well-placed cues improve retrieval accuracy by roughly 30–60 percentage points in most models, although the effect depends on model and placement. Large gains also occur for importance, and no affective family consistently dominates it, indicating a shared textual-salience effect across affectively framed and non-valenced cues. The gains generally do not broadly impair unmarked or updated memories. Internally, target–distractor selection emerges rapidly in mid-to-late layers, where cue-induced differences are strongest; attention shifts toward target values, and MLP interventions at selected layers increase target-value generation probability. Textual salience cues thus act as memory-priority signals that reshape selection among competing memories.

1 INTRODUCTION

As context windows expand, LLM agents can retain increasingly long interaction histories, including past experiences, user preferences, and intermediate decisions. Yet information being present in context does not guarantee reliable access when needed. For agents that reuse long-term histories, the challenge is therefore not only how much can be stored, but whether task-relevant memories can be accurately retrieved. Prior work shows that retrieval remains sensitive to position, context length, and interference even when the relevant information is present (Liu et al., 2024; Bai et al., 2024; Hsieh et al., 2024a; Kuratov et al., 2024).

Existing studies further show that models often retrieve beginning or end information more reliably than middle content (Liu et al., 2024), and that retrieval degrades when queries lack lexical overlap with the relevant text (Modarressi et al., 2025). These findings suggest that retrieval depends not only on whether information is stored, but also on how it is presented and competes with surrounding content. This raises a less studied question: when multiple memories coexist and compete, can a textual salience cue change their relative retrieval priority without changing the memory itself?

Emotional framing provides a natural motivation for studying textual salience. In human memory, arousal and valence influence memory processing (Cahill & McGaugh, 1995; Kensinger & Corkin, 2004; Mather & Sutherland, 2011), while recent work shows that emotionally framed prompts can alter LLM behavior and internal representations (Li et al., 2024; Reichman et al., 2026; Sofroniew et al., 2026; Sun et al., 2026). These observations motivate a broader question: can a short textual cue that signals salience or priority change the probability that an otherwise identical memory is retrieved from long context? We operationalize this question with four marker families. Threat,

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.

reward, and empathy provide affectively framed cues, while importance provides a non-valenced salience cue. We use these surface framings as controlled textual interventions and do not assume that they correspond to internal emotional states.

To isolate this effect from semantic knowledge and memory content, we construct a controlled long-context retrieval task with randomly assigned key–value bindings and similar-key distractors paired with different values. The random bindings force retrieval from context, while the distractors create competing memories around the target. For each instance, strictly paired conditions preserve the memory contents, target binding, and query, varying only the presence and placement of a short salience marker. Across five open-weight LLMs, most errors under similar-key interference match distractor values, indicating confusion among competing bindings rather than simple forgetting. Appropriately placed markers improve retrieval by roughly 30–60 percentage points in most models, although the effect depends on model and placement, and the gains do not broadly impair retrieval of other or updated memories. Crucially, the non-valenced importance cue is also highly effective, supporting a broader textual-salience interpretation of the shared behavioral effect. Internally, Logit Lens and Jacobian Lens reveal weak value-specific discrimination in early-to-middle layers and rapid target–distractor separation in mid-to-late layers, where cue–baseline differences are strongest. Attention analysis and MLP interventions further show that, in most models, target-oriented changes emerge at this stage and can partially shift retrieval toward the target. Together, these results suggest that salience markers modulate long-context retrieval by changing how competing memories are resolved into a final selection.

Our contributions are threefold. (1) We introduce a controlled framework that combines random key–value bindings, similar-key distractors, and strictly paired interventions to study how textual salience cues modulate long-context retrieval. (2) We characterize the behavioral effects and boundaries of four salience-marker families, showing substantial but model- and position-dependent changes in competitive retrieval; importantly, the non-valenced importance cue produces gains comparable to or larger than the affectively framed cues in several models, while the gains do not broadly degrade other memories under our conditions. (3) We connect these behavioral effects to internal target–distractor selection: Logit Lens, Jacobian Lens, attention analysis, and cross-sample MLP steering consistently place the strongest cue-related differences in mid-to-late layers, where targeted interventions can partially reproduce the behavioral direction.

2 Experimental Framework

To study whether textual salience cues can alter memory selection in long contexts, we need an experimental setting that separates retrieval itself from semantic knowledge and model priors. We therefore construct a controlled key–value retrieval framework in which competing memories are deliberately introduced, and paired interventions are used to examine how a salience marker changes retrieval of the target memory. Figure 1 illustrates the overall framework. This framework serves as the common basis for the behavioral experiments in Section 3 and the internal mechanistic analyses in Section 4.

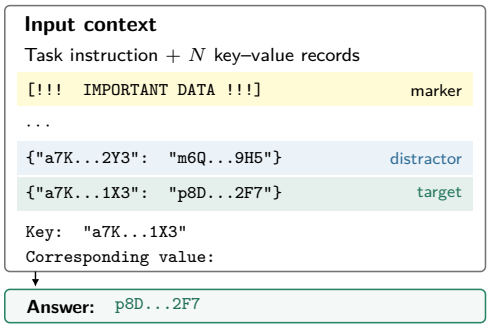


Figure 1: Controlled retrieval under memory competition. Similar-key distractors compete with the queried target binding, while a salience marker serves as a controlled textual intervention.

2.1 Retrieval under Memory Competition

Each example contains an ordered context $R = ((k_1, v_1), \ldots, (k_N, v_N))$ and a query key k^*. The model must identify the unique target binding (k^*, v^*) and generate v^*. Keys and values are randomly associated, so the answer must be retrieved from the current context rather than inferred from world knowledge or semantic associations.

We further introduce *similar-key distractors*: keys resembling k^* but paired with different values. These distractors create competing bindings, allowing us to distinguish simple retrieval failure from selection of an incorrect alternative. Construction and evaluation details are provided in Appendix C.

2.2 Textual Salience as a Controlled Intervention

Within this competitive setting, we introduce four salience-marker families—*threat*, *reward*, *empathy*, and *importance*—as short textual interventions. Threat uses negative/urgent framing, reward uses incentive framing, empathy uses interpersonal-affective framing, and importance is a non-valenced salience cue. These labels describe the surface framing of the text rather than internal states of the model; exact strings are listed in Appendix A.

We use a strict paired-intervention design. For each retrieval instance, the baseline and marked conditions share identical memory contents, target binding, and query. The baseline includes a matched blank placeholder at the same location, and downstream token positions are verified to remain aligned after tokenization. Thus, the comparison isolates the textual content introduced by the marker. Marker and target positions are varied independently to test whether the effect depends on their spatial relationship in the context. Specific configurations are given with the corresponding behavioral experiments.

2.3 Models and Evaluation

We evaluate five open-weight LLMs: Qwen3-8B (Yang et al., 2025), Qwen3.5-9B (Qwen Team, 2026a), Qwen3.6-35B-A3B (Qwen Team, 2026b), Gemma4-26B-A4B (Gemma Team, 2026), and Ministral-3-8B (Liu et al., 2026a), spanning different scales and both dense and mixture-of-experts architectures. Behavioral retrieval is evaluated by exact generation of the target value. For quantized checkpoints, mechanistic analyses use corresponding BF16 dequantized proxies, which we verify preserve the original models' retrieval behavior under matched conditions. Full model configurations, proxy construction, and validation are provided in Appendix B, with decoding and evaluation details in Appendix C. We first characterize retrieval failures under memory competition, then test whether salience markers alter this competition, and finally examine potential costs to accessing other memories.

3 Textual Salience Modulates Competitive Memory Retrieval

In Section 2, we constructed a controlled retrieval setting in which the target memory must compete with multiple similar bindings. In this section, we first use this framework to characterize the nature of long-context retrieval failures, then examine whether salience markers can alter this competition, and finally test whether the resulting retrieval gains come at the cost of accessing other memories. All comparisons use a paired design; complete experimental configurations and decoding details are provided in Appendix C.

3.1 Retrieval Failures Reflect Competition among Bindings

Before testing an intervention, we need to understand what it must change. Does interference lead to arbitrary answers, or to selection of a competing memory? We vary the number of similar-key distractors and the target's context position, recording retrieval accuracy and error destinations. Because distractor keys resemble the queried key but carry different random values, a distractor-value response identifies a specific competing binding. Complete settings, distractor sweeps, and cross-model results are provided in Appendix D.1.

Interference substantially impairs retrieval, although the response is position dependent and not strictly monotonic (Figure 2a). With 20 distractors, Qwen3-8B retrieves a beginning target with 98.5% accuracy, compared with 53.0% and 58.0% for middle and final targets. The target remains in the context in every case; what changes is the difficulty of selecting it among similar alternatives.

More diagnostic is where the errors land. At the strongest tested interference level, over 90% of Qwen3-8B errors at every target position are distractor-value substitutions. In the separate cross-model audit, 97.4–100% of errors also match distractor values (Figure 2b,c). The failure pattern is

therefore structured: the model typically returns a value associated with a competing key rather than an arbitrary out-of-context answer. This motivates the next test: can a textual salience cue shift the competition toward the target without changing the memory itself?

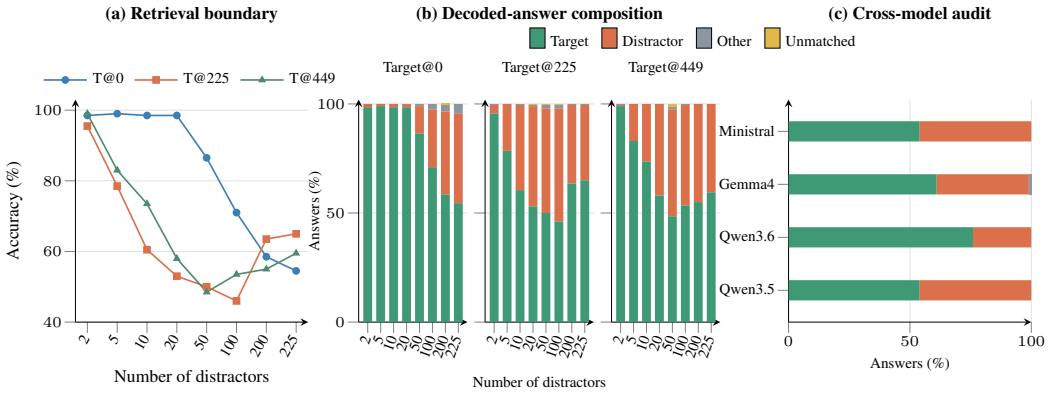


Figure 2: Retrieval failures predominantly select competing values. (a) Qwen3-8B accuracy across distractor counts. (b) Answer destinations in the same 450-record sweep. (c) Middle-target error composition for the other four models in a separate 1,000-record evaluation. Distractor substitutions dominate errors across models.

3.2 Salience Markers Shift Competitive Retrieval

We next ask whether salience markers can alter the competitive retrieval process identified above. Under substantial distractor interference, we test four marker families—*threat*, *reward*, *empathy*, and *importance*—while independently varying marker location and target position. The first three are affectively framed, whereas importance is a non-valenced salience cue. This design tests both where marker effects occur and whether large changes require affective framing. For clarity, Table 1 presents the complete importance results, while results for the other three families are reported in Appendix D.2.

Table 1: Importance-marker effects across models and positions. Entries give accuracy (%) and change from the no-marker baseline (percentage points). Both accuracy and gain are bold for positive changes. Qwen3-8B uses 200 examples per condition; the other models use 100.

Model	Target position	No marker (%)	Importance marker location: Start		Middle		End	
Qwen3-8B	Start	54.0	**66.5**	**(+12.5)**	**54.5**	**(+0.5)**	44.0	(−10.0)
	Middle	63.0	**65.5**	**(+2.5)**	**99.5**	**(+36.5)**	**69.0**	**(+6.0)**
	End	58.0	**67.0**	**(+9.0)**	**100.0**	**(+42.0)**	**65.0**	**(+7.0)**
Qwen3.5-9B	Start	95.0	**98.0**	**(+3.0)**	**99.0**	**(+4.0)**	**96.0**	**(+1.0)**
	Middle	56.0	56.0	(0.0)	**88.0**	**(+32.0)**	**58.0**	**(+2.0)**
	End	66.0	60.0	(−6.0)	**100.0**	**(+34.0)**	60.0	(−6.0)
Qwen3.6-35B-A3B	Start	100.0	100.0	(0.0)	100.0	(0.0)	100.0	(0.0)
	Middle	65.0	**68.0**	**(+3.0)**	**96.0**	**(+31.0)**	63.0	(−2.0)
	End	40.0	**42.0**	**(+2.0)**	**100.0**	**(+60.0)**	**51.0**	**(+11.0)**
Gemma4-26B-A4B	Start	89.0	85.0	(−4.0)	77.0	(−12.0)	74.0	(−15.0)
	Middle	62.0	**65.0**	**(+3.0)**	40.0	(−22.0)	**69.0**	**(+7.0)**
	End	94.0	90.0	(−4.0)	60.0	(−34.0)	93.0	(−1.0)
Ministral-3-8B	Start	89.0	87.0	(−2.0)	87.0	(−2.0)	81.0	(−8.0)
	Middle	29.0	**35.0**	**(+6.0)**	**74.0**	**(+45.0)**	29.0	(0.0)
	End	53.0	**56.0**	**(+3.0)**	**100.0**	**(+47.0)**	**72.0**	**(+19.0)**

Short salience markers can substantially alter long-context retrieval, with strong effects in four of the five models. Table 1 shows this pattern for the importance marker: for a middle target, accuracy

rises from 63.0% to 99.5% in Qwen3-8B, from 56% to 88% in Qwen3.5-9B, from 65% to 96% in Qwen3.6-35B-A3B, and from 29% to 74% in Ministral-3-8B. The complete four-family results show that large gains do not require affective framing: under the same middle-marker/middle-target condition, importance produces the largest gain for Qwen3-8B and Qwen3.5, whereas empathy is larger for Qwen3.6 and Ministral; Gemma4 decreases under all four families (Appendix D.2). Thus, the shared behavioral pattern is best characterized as textual-salience modulation, with effect direction and magnitude varying across models, marker families, and placements.

Marker placement is also crucial. Middle markers generally produce stronger retrieval changes than start or end markers, suggesting that salience markers modulate the priority of particular memories rather than uniformly improving retrieval across the context. The effect is also model dependent: in Gemma4-26B-A4B, all four middle-position marker families reduce middle-target accuracy from 62% to 32–46% (Appendix D.2). Overall, salience markers systematically reshape competitive retrieval, with effects depending on the model, marker family, and context position.

Do additional placements help? The strong placement dependence raises a natural question: can multiple markers reinforce one another, or does their arrangement matter more than their number? We repeat each marker at two or all three locations and compare each combination with that family's best single placement.

For Qwen3-8B middle targets, Figure 3 shows that combinations involving Middle are generally strongest, whereas Start+End is much weaker. Start+Middle outperforms the best single placement for threat, reward, and empathy, while importance is already near ceiling with a single middle marker.

Change from best single (pp)

	S+M	S+E	M+E	All
Threat	+18.5	-7.5	+14	+13.5
Reward	+24	+0.5	+13	+14.5
Empathy	+6	-29	+4	+3.5
Importance	-3	-24.5	+0.5	0
Mean	**+11.4**	**-15.1**	**+7.9**	**+7.9**

Figure 3: Qwen3-8B, middle target. Each entry is the accuracy of a position combination minus that family's best single-placement accuracy (percentage points). Mean averages across marker families; positive values favor the combination. S/M/E: start/middle/end.

Cross-model results show the same broader pattern: combinations involving Middle can provide additional gains, whereas Start+End is typically weaker. The benefit can also saturate or reverse, with Gemma4 again showing a distinct pattern. Thus, performance depends more on which locations are combined than on marker count, with Middle playing a central role. Complete settings and cross-model results are in Appendix D.3.

3.3 Retrieval Gains Are Largely Selective

If a salience marker works mainly by concentrating the model's attention on a particular region, then enhancing one memory may come at the expense of others. We therefore ask whether the resulting retrieval change is selective, or whether it broadly disrupts access to other information in the context.

Does enhancement impair access to other memories? We fix the marker's location and vary which context record is queried. If enhancement broadly diverts retrieval away from the rest of the context, other targets should deteriorate. In Qwen3-8B, a middle marker raises accuracy at the aligned target from 68.0% to 100% with importance and to 95.0% with empathy (Figure 4a). Changes elsewhere are smaller and mixed in direction, without a context-wide decline. The contrast between the aligned target and other locations indicates a selective change in access rather than a uniform shift across all records. Other marker placements and models also show position-specific costs, but the large aligned-target gains are not generally accompanied by broad degradation. Complete cross-model and multi-position results are reported in Appendix D.4.

Does a marked memory resist updating? A stricter test asks whether prioritizing an old binding prevents the model from using a newer one. We update the same key once or twice, always query its latest value, and compare marking the old binding, marking the latest binding, and no marker. This creates a direct conflict between the previously marked information and the value required by the current query.

(a) Retrieval of unmarked targets: marker at middle

Baseline (%)	94.5	90	81	78	66	62.5	68.5	68	64	65.5	58	68.5	72	85.5	69.5
Importance Δ	-3	0	0	+2.5	+4	+6.5	+7	+32	+10.5	+9	+4	+0.5	+2	0	+5
Empathy Δ	-1	+2.5	+3	+0.5	+5.5	+7.5	+7	+27	+5.5	+4	+2.5	+2	-0.5	+0.5	+6.5
	-200	-150	-100	-50	-20	-10	-5	**0**	+5	+10	+20	+50	+100	+150	+200

Target position relative to marker (equally spaced tested positions)

(b) Memory updating

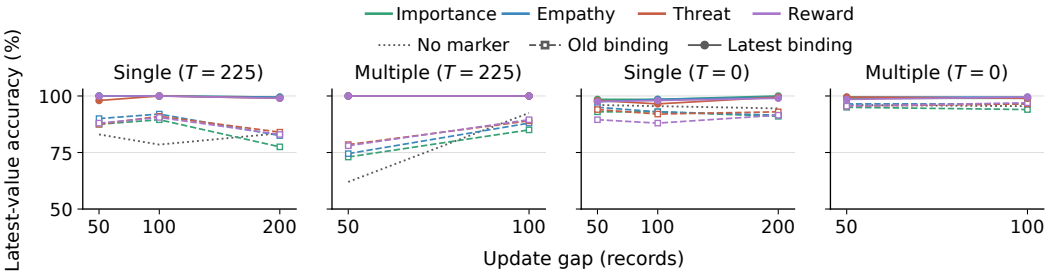


Figure 4: Selectivity of retrieval modulation in Qwen3-8B. (a) Fixed middle marker: baseline accuracy (%) and marker-induced changes (percentage points), with the aligned target boxed. Columns show discrete tested positions. (b) Latest-value retrieval after one or two updates, comparing no marker, an old-binding marker, and a latest-binding marker.

In Qwen3-8B, marking the latest binding preserves 96.5–100% latest-value accuracy across tested settings (Figure 4b). Marking the old binding also causes no broad loss, although individual conditions change. Cross-model results further show that update interference can make the unmarked task difficult: some no-marker conditions reach zero accuracy, while marked conditions recover access to the latest value. Full family-, model-, and gap-specific results are in Appendix D.5.

Taken together, these results show that salience markers do not simply make all information in the context easier to retrieve. Their effects emerge primarily under competitive retrieval, depend strongly on marker placement, and are largely concentrated on the associated target memory rather than producing a uniform improvement across the context.

4 How Salience Markers Reshape Retrieval Selection

The behavioral results in Section 3 show that long-context retrieval failures largely reflect incorrect selection between target and competing memories, and that salience markers can substantially alter this competition. This raises a natural question: when does this change emerge inside the model, and how does it influence the final retrieval decision?

We address this question from three complementary perspectives. First, we trace when target and distractor values separate across layers. Second, we examine whether a salience marker changes access to target and distractor memories. Finally, we intervene on internal representations to test whether marker-induced changes can directly shift retrieval toward the target. We focus on the middle-marker/middle-target importance condition with Qwen3-8B as the main case. Its strong behavioral effect makes it a direct probe of the broader salience effect. Analysis protocols are given in Appendix E, with cross-model results in Appendix F.

4.1 Target–Distractor Selection Forms in Mid-to-Late Layers

The behavioral results show different final selections between target and distractor memories, but not when this separation forms. We therefore apply Logit Lens (nostalgebraist, 2020) and Jacobian Lens (Gurnee et al., 2026) layer by layer immediately before the first value token, comparing support for the target, distractors, and control candidates. Candidate construction, alignment, and readout details are given in Appendix E.2.

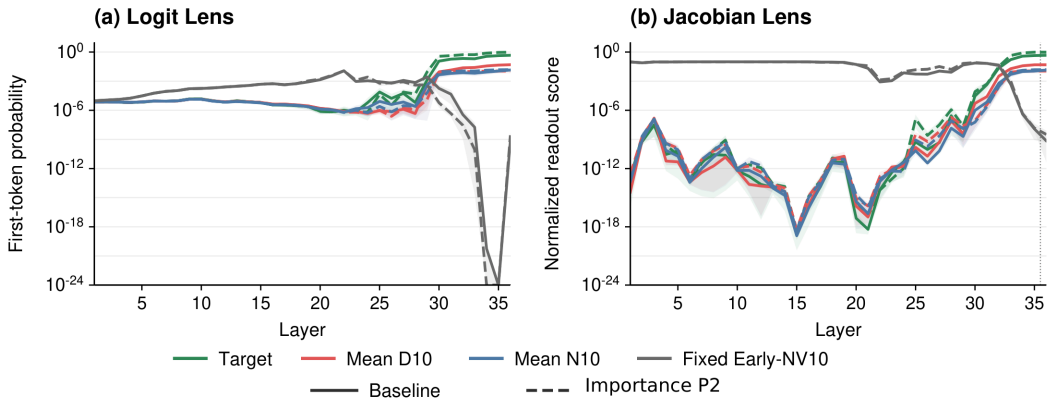


Figure 5: Emergence of retrieval selection in Qwen3-8B. Colors identify fixed candidate groups; solid/dashed lines denote baseline/importance. Shading shows bootstrap 95% confidence intervals. The left panel gives Logit Lens probabilities; the right gives normalized Jacobian Lens readout scores.

Figure 5 reveals a two-stage pattern. In early-to-middle layers, candidate values remain weakly separated; in mid-to-late layers, target–distractor separation grows and the importance–baseline difference becomes strongest. At the final Qwen3-8B LM head, importance raises the mean target first-token probability from 0.4674 to 0.9743 while reducing mean distractor (D10) probability from 0.0497 to 0.0110. The marker-induced behavioral difference becomes prominent as target–distractor selection emerges, rather than being fixed in early layers. Other models show similar late separation, with model-dependent marker effects. Complete cross-model results are in Appendix F.1.

These readouts answer the "when" question: the marker's effect is strongest in mid-to-late layers as competing memories are resolved into a retrieval choice. We next ask whether this shift is accompanied by changes in access to target and distractor memories.

4.2 Attention Shifts toward Target Memories

If a salience marker alters target–distractor selection in mid-to-late layers, one possible mechanism is that the model draws more information from the target memory and less from competing memories. We test this by comparing baseline and importance conditions at the answer-prediction position, measuring attention to target and distractor bindings. Source-span definitions, normalization, and aggregation details are given in Appendix E.3.

Figure 6 shows that this attention shift is also concentrated in mid-to-late layers. In Qwen3-8B, importance increases attention to the target value while reducing attention to distractor values. The increase peaks at L25 at 4.94×10^{-3} per source token, while distractor-value attention declines across L10–L36. This aligns with Section 4.1: as target–distractor separation emerges, information flow to the answer position shifts toward the target memory.

The other Qwen models and Ministral show similar target-oriented patterns, whereas Gemma4 shows the opposite late-layer trend, consistent with its behavioral reversal. Complete cross-model results are reported in Appendix F.2. To complement this observational evidence, we next test whether marker-induced internal representations can directly shift retrieval toward the target.

4.3 Shared MLP Changes Can Influence Held-Out Retrieval

The previous analyses show that importance-related behavioral changes coincide with mid-to-late-layer target–distractor separation and attention redistribution. We therefore ask whether injecting an importance-induced internal change into the baseline can directly shift retrieval toward the target.

We derive layer-wise MLP directions from paired baseline–importance differences and use held-out cross-fitting so that each test example is steered only with directions estimated from other examples. Each direction is injected into the baseline at one layer, before residual addition at the retrieval

	Target key	Target value	D key	D value	Final query key		Target key	Target value	D key	D value	Final query key
L1	+1.65e-05	+5.02e-05	+0.00016	+0.000124	+0.0792	L19	+0.628	+0.299	-0.00103	-0.000648	-0.661
L2	+0.000327	+8.34e-05	+0.000237	+0.000154	+0.0187	L20	+0.258	+0.43	+0.00147	-0.00112	-0.762
L3	+0.00743	+0.00814	+0.000154	+4.47e-05	+0.0269	L21	+0.46	+2.91	+0.00167	-0.00823	-0.326
L4	+0.00875	+0.00144	-1.45e-05	+8.86e-06	-0.00532	L22	+0.642	+2.66	-0.00411	-0.0133	-1.23
L5	+0.012	+0.00112	+9.82e-05	+5.62e-05	-0.00786	L23	+0.795	+1.02	-0.000386	-0.00629	-2.21
L6	+0.0151	+0.00154	+0.000109	+0.000108	+0.0132	L24	+0.768	+4.18	-0.00811	-0.0295	-1.97
L7	+0.0245	+0.00422	-8.51e-06	+2.12e-05	-0.0193	L25	+1.02	+4.94	-0.00859	-0.0391	-0.862
L8	+0.0509	+0.00865	+8.19e-05	+2.91e-05	-0.0102	L26	+0.0168	+0.767	+0.00541	-0.00357	+0.0845
L9	+0.0566	+0.0166	+2.36e-05	+1.49e-05	-0.0587	L27	+0.238	+1.06	+0.00196	-0.0261	-0.139
L10	+0.0615	+0.0206	-4.16e-05	-5.96e-05	-0.106	L28	+0.172	+0.584	+0.00342	-0.013	+0.187
L11	+0.0342	+0.027	-0.000421	-0.000213	-0.0628	L29	+0.0245	+1.95	+0.000592	-0.0211	+0.0458
L12	+0.0476	+0.0292	-0.000179	-0.000235	-0.0722	L30	+0.062	+1.92	+0.00415	-0.0465	+0.0907
L13	+0.0577	+0.0673	-2.73e-06	-0.00159	-0.167	L31	-0.00269	+0.281	+0.00145	-0.0174	+0.0535
L14	+0.211	+0.167	-0.000527	-0.000391	-0.324	L32	+0.12	+1.97	-0.00194	-0.0207	+0.206
L15	+0.0807	+0.0424	-0.000514	-0.000319	-0.46	L33	+0.0727	+0.884	-0.00154	-0.0197	+0.265
L16	+0.101	+0.0511	-0.00101	-0.000685	-0.0977	L34	+0.198	+1.58	-0.00687	-0.0344	+0.236
L17	+0.306	+0.346	-0.00142	-0.00124	-0.251	L35	+0.136	+0.616	-0.00454	-0.0105	+0.374
L18	+0.451	+0.459	-0.000495	-0.000481	-0.233	L36	+0.0396	+0.014	+0.000329	-0.00028	-0.0882

Importance - baseline

-5 0 5 Per source token (x 10^-3)

Figure 6: Memory-source attention during retrieval in Qwen3-8B. Cells give importance minus baseline attention per source token, averaged over heads, in units of 10^{-3}. Positive values indicate increased attention under importance.

position, and we measure the resulting change in target first-token probability. MLP-output cosine similarity likewise shows the largest directional divergence in later layers. Full construction and intervention details are provided in Appendix E.5, with cosine results in Appendix F.3.

Figure 7 shows that positive steering effects concentrate at selected mid-to-late layers. In Qwen3-8B, interventions at L28 and L29 increase target first-token probability by 0.0200 and 0.0183 on average. The other Qwen models and Ministral show related effects, while Gemma4 is smaller and mixed. Complete cross-model results are in Appendix F.4. The held-out transfer shows that these importance-induced representations can directly influence target selection, although single-layer steering only partially reproduces the full behavioral effect.

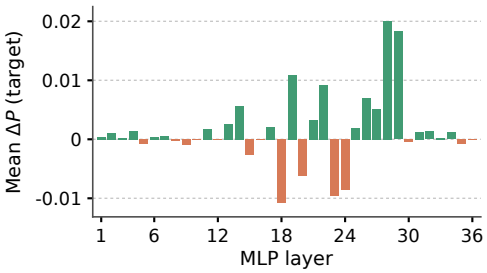


Figure 7: Qwen3-8B shared MLP steering. Bars show mean changes in target first-token probability after held-out, single-layer interventions. Green/orange denotes increases/decreases.

Together, the three analyses support a consistent picture: importance-related differences become prominent as the model enters the mid-to-late-layer target–distractor selection stage, alongside a shift in information access toward the target. Intervening on MLP representations at these layers can partially shift retrieval in the same direction. Thus, this mechanistic case study shows how a textual salience marker can modulate the stage at which competing memories are resolved into the final retrieval decision, rather than acting as a uniform enhancement signal throughout the network.

5 Related Work

Long-Context Retrieval, Memory, and Textual Salience in LLMs. Long-context research shows that larger context windows do not guarantee reliable retrieval. Benchmarks such as LongBench, RULER, and BABILong evaluate retrieval and reasoning over extended contexts (Bai et al., 2024; Hsieh et al., 2024a; Kuratov et al., 2024), while studies of positional sensitivity, attention calibration, lexical matching, and distractor interference reveal strong dependence on where and

how information appears (Liu et al., 2024; Hsieh et al., 2024b; Modarressi et al., 2025; Shi et al., 2023). Retrieval-augmented and long-term memory models extend access beyond the immediate context (Lewis et al., 2020; Borgeaud et al., 2022; Wang et al., 2023), and agent-memory systems and benchmarks study memory over extended interaction histories (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024; Maharana et al., 2024; Wu et al., 2025a). In parallel, work on LLM emotion examines emotional prompting, capabilities, and internal representations (Li et al., 2024; Sabour et al., 2024; Reichman et al., 2026; Sofroniew et al., 2026; Sun et al., 2026), including recent studies of long interactions and memory-oriented agents (Liu et al., 2025; Fu et al., 2026; Liu et al., 2026b; Amanlou et al., 2026). Emotional framing is one source of textual salience, while our intervention set also includes a non-valenced importance cue. We connect these lines of work by asking whether short textual salience cues can shift retrieval priority among competing memories while leaving memory content fixed.

Mechanistic Analysis of LLM Retrieval. Mechanistic interpretability offers several tools for tracing retrieval through a model. Logit Lens, Tuned Lens, and related vocabulary-space analyses probe intermediate representations (nostalgebraist, 2020; Belrose et al., 2023; Dar et al., 2023; Geva et al., 2022), while Jacobian Lens accounts for downstream transformations (Gurnee et al., 2026). Studies of induction and retrieval heads examine contextual access through attention (Olsson et al., 2022; Wu et al., 2025b), and factual-recall work traces how stored information becomes output predictions (Geva et al., 2023). Complementary interventions—including causal tracing, Activation Addition, Contrastive Activation Addition, representation engineering, and activation patching—test whether internal representations or components influence model outputs (Meng et al., 2022; Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023; Zhang & Nanda, 2024). Building on these approaches, we combine layer-wise readouts, source-level attention analysis, and held-out MLP interventions to characterize salience-modulated target–distractor selection.

6 Limitations

First, our controlled key–value setting isolates the effect of textual salience cues on competitive retrieval, but natural agent histories contain richer semantic structure shaped by relevance, recency, and user importance. Second, the marker strings vary along multiple surface-semantic dimensions. Threat, reward, and empathy are affectively framed, whereas importance is non-valenced; their shared effects support a general salience interpretation, but do not isolate the contributions of lexical content, urgency, importance, interpersonal framing, or punctuation. Tighter controls, factorial cue manipulations, and independent ratings could further disentangle these factors, while Gemma4's reversed effect highlights model dependence. Third, although layer-wise readouts, attention analysis, and MLP interventions consistently implicate mid-to-late-layer target–distractor selection, finer component-level ablations, activation patching, and cross-model transfer are needed to identify which internal components are necessary or sufficient for the effect.

7 Conclusion

We studied whether lightweight textual salience cues can change which memory an LLM selects when multiple memories compete. Controlled retrieval experiments show that many failures reflect incorrect selection among competing bindings rather than simple forgetting. Under such conditions, salience markers can substantially alter target retrieval, with effects depending strongly on model and placement. The effect spans affectively framed cues and a non-valenced importance cue, with no affective family consistently producing the largest change. Our internal analyses further show that these differences become most pronounced during mid-to-late-layer target–distractor selection, accompanied by systematic changes in attention and internal representations.

These results show that long-context retrieval depends not only on what information is stored, but also on how models prioritize competing memories during selection. Textual salience cues provide a controlled window into this process: identical information can receive different retrieval priority depending on surrounding textual cues. These findings offer a new perspective on memory selection in long-context LLMs and agents, suggesting that future memory systems should consider not only what information is stored, but also how retrieval priority is determined.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing, translation, and figure formatting. All AI-assisted content was reviewed and verified by the authors against the underlying experiments and source materials. The authors take full responsibility for the final content of this work, including its text, claims, analyses, and artifacts.

REPRODUCIBILITY STATEMENT

Code and experiment files for reproducing the main behavioral and mechanistic analyses are available in our anonymous repository: `TEXTUAL-SALIENCE-55E1`. Additional experimental configurations, implementation details, and complete results are provided in the appendix.

REFERENCES

Mohammad Amanlou, Parham Abed Azad, Farbod Davoodi, Mostafa Masumi, Behnam Bahrak, and Abdol-Hossein Vahabie. PsychoAgent: An affect-sensitive cognitive architecture for conflict-aware memory in LLM agents. arXiv preprint arXiv:2608.07438, 2026. URL https://arxiv.org/abs/2608.07438.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112, 2023. URL https://arxiv.org/abs/2303.08112.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2206–2240, 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

Larry Cahill and James L. McGaugh. A novel demonstration of enhanced memory associated with emotional arousal. *Consciousness and Cognition*, 4(4):410–421, 1995. doi: 10.1006/ccog.1995.1048. URL https://pubmed.ncbi.nlm.nih.gov/8750416/.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170, 2023. doi: 10.18653/v1/2023.acl-long.893. URL https://aclanthology.org/2023.acl-long.893/.

Xing Fu, Yulin Hu, Mengtong Ji, Haozhen Li, Yixin Sun, Weixiang Zhao, Yanyan Zhao, and Bing Qin. ENPMR-Bench: Benchmarking proactive memory retrieval for emotional support agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 41910–41933, 2026. doi: 10.18653/v1/2026.findings-acl.2080. URL https://aclanthology.org/2026.findings-acl.2080/.

Gemma Team. Gemma 4 technical report. arXiv preprint arXiv:2607.02770, 2026. URL https://arxiv.org/abs/2607.02770.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, 2022. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. arXiv preprint arXiv:2607.15495, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024a. URL https://arxiv.org/abs/2404.06654.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14982–14995, 2024b. doi: 10.18653/v1/2024.findings-acl.890. URL https://aclanthology.org/2024.findings-acl.890/.

Elizabeth A. Kensinger and Suzanne Corkin. Two routes to emotional memory: Distinct neural processes for valence and arousal. *Proceedings of the National Academy of Sciences*, 101(9): 3310–3315, 2004. doi: 10.1073/pnas.0306408101. URL https://doi.org/10.1073/pnas.0306408101.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.10149.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://arxiv.org/abs/2005.11401.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. The good, the bad, and why: Unveiling emotions in generative AI. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28905–28934, 2024. URL https://proceedings.mlr.press/v235/li24bs.html.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, et al. Ministral 3. arXiv preprint arXiv:2601.08584, 2026a. URL https://arxiv.org/abs/2601.08584.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Peng Liu, Zhen Tao, Jihao Zhao, Ding Chen, Yansong Zhang, Cuiping Li, Zhiyu Li, and Hong Chen. MemEmo: Evaluating emotion in memory systems of agents. arXiv preprint arXiv:2602.23944, 2026b. URL https://arxiv.org/abs/2602.23944.

Weichu Liu, Jing Xiong, Yuxuan Hu, Zixuan Li, Minghuan Tan, Ningning Mao, Hui Shen, Wendong Xu, Chaofan Tao, Min Yang, Chengming Li, Lingpeng Kong, and Ngai Wong. LongEmotion: Measuring emotional intelligence of large language models in long-context interaction. arXiv preprint arXiv:2509.07403, 2025. URL https://arxiv.org/abs/2509.07403. Version 2, revised January 2026.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Mara Mather and Matthew R. Sutherland. Arousal-biased competition in perception and memory. *Perspectives on Psychological Science*, 6(2):114–133, 2011. doi: 10.1177/1745691611400234. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3110019/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843, 2016. URL https://arxiv.org/abs/1609.07843.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-context evaluation beyond literal matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44554–44570, 2025. URL https://proceedings.mlr.press/v267/modarressi25a.html.

nostalgebraist. Interpreting GPT: The logit lens. AI Alignment Forum, 2020. URL https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. arXiv preprint arXiv:2209.11895, 2022. URL https://arxiv.org/abs/2209.11895.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. arXiv preprint arXiv:2310.08560, 2023. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763. URL https://arxiv.org/abs/2304.03442.

Qwen Team. Qwen3.5-9B: Official model card. Hugging Face model repository, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Qwen Team. Qwen3.6-35B-A3B: Official model card. Hugging Face model repository, 2026b. URL https://huggingface.co/Qwen/Qwen3.6-35B-A3B.

Benjamin Reichman, Adar Avasian, Samuel Webster, and Larry Heck. Emotion is not just a label: Latent emotional factors in LLM processing. arXiv preprint arXiv:2603.09205, 2026. URL https://arxiv.org/abs/2603.09205.

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

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5986–6004, 2024. doi: 10.18653/v1/2024.acl-long.326. URL https://aclanthology.org/2024.acl-long.326/.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. arXiv preprint arXiv:2604.07729, 2026. URL https://arxiv.org/abs/2604.07729.

Moran Sun, Tianlin Li, Yuwei Zheng, Zhenhong Zhou, Aishan Liu, Xianglong Liu, and Yang Liu. How emotion shapes the behavior of LLMs and agents: A mechanistic study. arXiv preprint arXiv:2604.00005, 2026. URL https://arxiv.org/abs/2604.00005.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. arXiv preprint arXiv:2308.10248v1, 2023. URL https://arxiv.org/abs/2308.10248v1.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper/2023/hash/ebd82705f44793b6f9ade5a669d0f0bf-Abstract-Conference.html.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9b77f07301b1ef1fe810aae96c12cb7b-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

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

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023. URL https://arxiv.org/abs/2310.01405.

A Textual Salience Marker Strings and Intervention Scope

Operational definition. The four families describe the textual framing of the intervention. Threat, reward, and empathy are affectively framed, while importance is a non-valenced salience cue. These are operational labels for the marker text, not claims about an internal emotional state of the model. Table 2 gives the exact strings inserted around the key–value context while keeping the target value unchanged.

Table 2: Exact marker strings used in the single-position, combination, unmarked-target, and memory-update experiments.

Family	Marker text
Threat	`[!!! CRITICAL WARNING - SYSTEM RECOVERY DEPENDS ON THIS !!!]`
Reward	`[!!! OPPORTUNITY - REMEMBER THIS FOR MAXIMUM REWARD !!!]`
Empathy	`[!!! PLEASE - THIS NEXT PART MEANS EVERYTHING TO ME !!!]`
Importance	`[!!! IMPORTANT DATA !!!]`

What is held fixed. Paired conditions share the ordered memory bindings, queried key, target value, target record position, and decoding configuration. Marker and target locations vary independently. The baseline contains a matched blank placeholder at every insertion site, with downstream token positions checked after tokenization. Markers leave memory records intact. Combination conditions repeat one family at the selected sites, without mixing families.

B Model Configuration and Proxy Validation

Analysis backends. Qwen3-8B and Qwen3.5-9B use their original weights. The other analyses use BF16 dequantized proxies: AWQ-derived for Qwen3.6-35B-A3B and Gemma4-26B-A4B, and FP8-derived for Ministral-3-8B. Architecture, tokenizer, and prompt protocol remain unchanged. MLP denotes the full feed-forward branch entering residual addition, including expert aggregation in MoE models. Table 3 summarizes the backends and feed-forward branches.

Table 3: Internal-analysis backends and feed-forward branches. Layer counts refer to the language-model stack; attention-layer coverage is specified in Appendix E.3.

Model	Layers	Internal-analysis backend and branch
Qwen3-8B	36	Original weights; dense MLP
Qwen3.5-9B	32	Original weights; dense MLP
Qwen3.6-35B-A3B	40	AWQ-dequantized BF16; routed MoE plus gated shared expert
Gemma4-26B-A4B	30	AWQ-dequantized BF16; combined dense FFN and routed MoE
Ministral-3-8B	34	FP8-dequantized BF16; dense MLP

Matched validation protocol. Each original–proxy pair uses the same 100 examples with 1,000 records, 10 similar-key distractors, and target index 500. The marked condition inserts Middle importance. Table 4 reports exact-match generated-value retrieval accuracy.

Interpretation. All three proxies preserve the marker effect's direction, with marker-induced changes differing by at most two percentage points. Both versions improve for Qwen3.6 and Ministral and decline for Gemma4, supporting the matched internal analyses.

Table 4: Original checkpoint versus BF16 proxy: retrieval accuracy (%) and marker-induced change (pp). These matched validation runs are separate from the single-position behavioral sweeps.

	Original checkpoint			BF16 proxy		
Model	Baseline	Importance	Δ	Baseline	Importance	Δ
Qwen3.6-35B-A3B	76	96	+20	76	98	+22
Gemma4-26B-A4B	62	40	−22	62	40	−22
Ministral-3-8B	54	72	+18	57	74	+17

C Behavioral Experimental Details

C.1 Context Construction and Paired Conditions

Each example consists of a JSON-style sequence of key–value records followed by a query of the form `Key: "`k^*`"` and a request for the `Corresponding value`. Keys and values are fixed-length random alphanumeric strings, typically 16 characters in the core setting. Values are assigned independently of key meaning, removing an informative semantic relationship between a key and its answer. The queried key identifies one target binding, except in the update experiment, where repeated bindings for that key define a temporal sequence.

Similar-key distractors perturb two characters of the target key and are paired with distinct non-target random values. The remaining records are unrelated background bindings. Changing the distractor count changes the composition of a context of fixed record count; it does not increase its number of records. All positions and gaps are specified in zero-based record indices; an insertion anchor may also lie after the final record.

Paired records and queries, and the independent control of marker and target positions, follow Appendix A. Combination and update manipulations are specified in their experiment subsections below.

C.2 Decoding, Matching, and Units

All behavioral experiments use greedy decoding with the model's native chat template and a direct value-retrieval instruction. Thinking mode is disabled for the Qwen models. The Qwen3-8B output parser removes an optional `<think>...</think>` span before extracting the value; Gemma4 and Ministral directly return values under their native templates.

Let $\hat{v}_i$ be the extracted generated value and v_i^* the requested value. Accuracy in percent is $A = 100n^{-1}\sum_i \mathbf{1}[\hat{v}_i = v_i^*]$. A difference between two such percentages is reported in *percentage points* (pp), not relative percent improvement. In the update experiment v_i^* is the latest value. Internal analyses measure support for the target value's first token, as defined in Appendix E.

For error analysis, answers are categorized as the target value, a distractor value, another value present in the context, or an unmatched answer. The distractor share among errors is the number of distractor substitutions divided by all incorrect answers; it differs from the distractor share among all outputs. Single-marker tables highlight positive changes from the paired baseline; the combination tables highlight the highest accuracy within each row.

C.3 Experiment Map and Sample Scope

Table 5 summarizes the five behavioral experiments and their diagnostic questions.

The different experiments are separate evaluations. For example, Qwen3-8B's middle-target baselines are 65.0% in the 225-distractor boundary sweep, 63.0% in the single-marker/combination reference, and 68.0% at the aligned target in the fixed-middle-marker sweep. Each marked condition is compared with the no-marker baseline from the same evaluation.

Table 5: Behavioral experiment map. Qwen3-8B uses 450 records and 200 examples per condition; the other models use 1,000 records and 100 examples.

Experiment	Manipulation	Diagnostic question
Distractor boundary	Distractor count and target position	Where do retrieval errors go?
Single marker	Family, marker location, target position	Can a cue shift competitive retrieval?
Position combination	Two or three sites for one family	Does repetition beat the best single site?
Unmarked targets	Fixed marker, variable queried record	Is enhancement paid for elsewhere?
Memory updating	Update gap; old or latest marking	Does priority produce stale retrieval?

D Additional Behavioral Results

D.1 Distractor Boundaries and Error Destinations

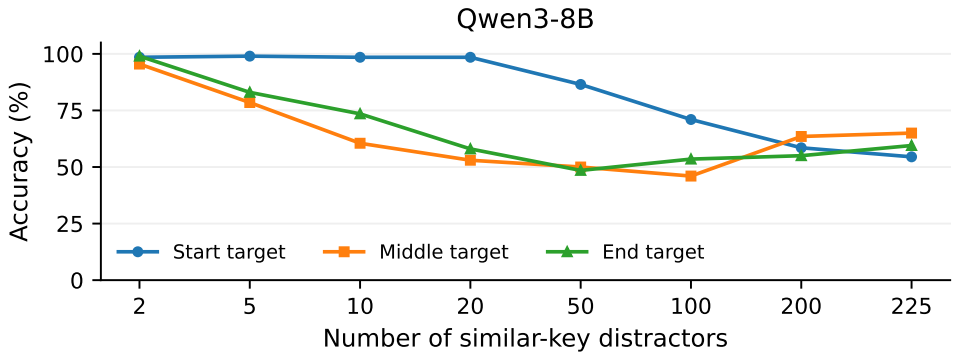
Purpose and protocol. The accuracy sweep establishes that similar bindings interfere with retrieval; the error audit then distinguishes confusion with a competing value from an answer outside the context. Qwen3-8B uses $N = 450$, $n = 200$, target indices 0/225/449, and distractor counts 2, 5, 10, 20, 50, 100, 200, and 225. The other models use $N = 1{,}000$, $n = 100$, target indices 0/500/999, and counts 2, 5, 10, 20, 50, 100, 200, and 500. Each context retains its fixed total record count. The complete curves are shown for Qwen3-8B (Figure 8), Qwen3.5-9B (Figure 9), Qwen3.6-35B-A3B (Figure 10), Gemma4-26B-A4B (Figure 11), and Ministral-3-8B (Figure 12).

Error composition. Table 6 gives all Qwen3-8B outcome counts. At 225 distractors, 82/91, 69/70, and 81/81 errors are distractor substitutions at Start, Middle, and End, respectively. The separate 100-example middle-target audit records target/distractor/other counts of 54/46/0 for Qwen3.5, 76/24/0 for Qwen3.6, 61/38/1 for Gemma4, and 54/46/0 for Ministral. The corresponding distractor shares among errors are 100%, 100%, 97.4%, and 100%. Across models, the same pattern emerges: incorrect answers overwhelmingly select a competing value already present in the context. This makes target–distractor selection a natural focus for the marker intervention.

Table 6: Complete Qwen3-8B boundary outcome counts ($n = 200$ per row). Other denotes a non-target, non-distractor context value; unmatched denotes no context-value match. D/errors is conditional on an incorrect answer.

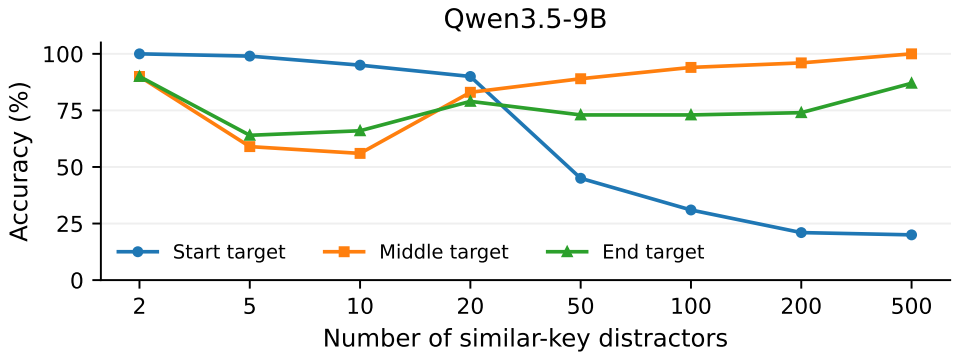
D	Target	Correct	D value	Other	Unmatched	D/errors (%)
2	Start	197	3	0	0	100.0
2	Middle	191	8	1	0	88.9
2	End	198	1	1	0	50.0
5	Start	198	2	0	0	100.0
5	Middle	157	43	0	0	100.0
5	End	166	34	0	0	100.0
10	Start	197	3	0	0	100.0
10	Middle	121	78	1	0	98.7
10	End	147	53	0	0	100.0
20	Start	197	3	0	0	100.0
20	Middle	106	92	1	1	97.9
20	End	116	84	0	0	100.0
50	Start	173	25	2	0	92.6
50	Middle	100	96	3	1	96.0
50	End	97	98	2	3	95.1
100	Start	142	53	5	0	91.4
100	Middle	92	104	3	1	96.3
100	End	107	92	1	0	98.9
200	Start	117	76	6	1	91.6
200	Middle	127	72	1	0	98.6
200	End	110	90	0	0	100.0
225	Start	109	82	9	0	90.1
225	Middle	130	69	1	0	98.6
225	End	119	81	0	0	100.0

Non-monotonic boundaries. Increasing distractor count produces different trajectories across target positions. Some middle and final targets recover at higher counts, whereas beginning targets can deteriorate. The effect of interference therefore depends on the target's context position as well as the number of competitors. The single-marker experiment uses configurations with substantial retrieval degradation identified in these sweeps.



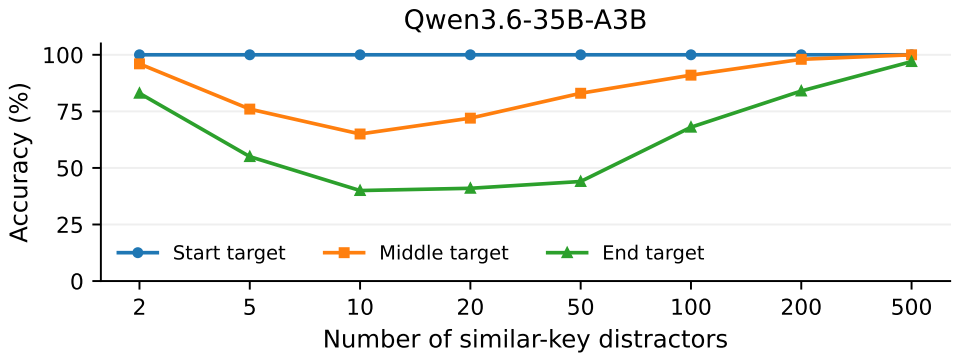
Target / D	2	5	10	20	50	100	200	225
Start	98.5	99.0	98.5	98.5	86.5	71.0	58.5	54.5
Middle	95.5	78.5	60.5	53.0	50.0	46.0	63.5	65.0
End	99.0	83.0	73.5	58.0	48.5	53.5	55.0	59.5

Figure 8: Complete distractor boundary for Qwen3-8B: all eight counts and three target positions. The numeric panel reports accuracy (%). There are 200 examples per condition in a 450-record context. Horizontal positions are discrete tested counts.



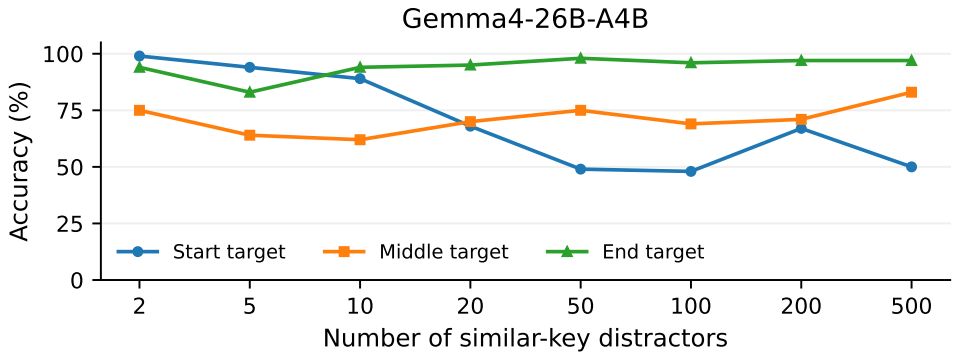
Target / D	2	5	10	20	50	100	200	500
Start	100.0	99.0	95.0	90.0	45.0	31.0	21.0	20.0
Middle	90.0	59.0	56.0	83.0	89.0	94.0	96.0	100.0
End	90.0	64.0	66.0	79.0	73.0	73.0	74.0	87.0

Figure 9: Complete distractor boundary for Qwen3.5-9B: all eight counts and three target positions. The numeric panel reports accuracy (%). There are 100 examples per condition in a 1,000-record context. Horizontal positions are discrete tested counts.



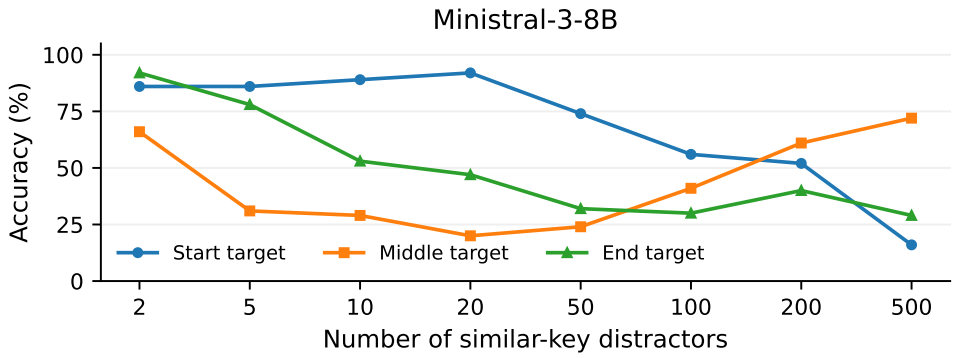
Target / D	2	5	10	20	50	100	200	500
Start	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Middle	96.0	76.0	65.0	72.0	83.0	91.0	98.0	100.0
End	83.0	55.0	40.0	41.0	44.0	68.0	84.0	97.0

Figure 10: Complete distractor boundary for Qwen3.6-35B-A3B: all eight counts and three target positions. The numeric panel reports accuracy (%). There are 100 examples per condition in a 1,000-record context. Horizontal positions are discrete tested counts.



Target / D	2	5	10	20	50	100	200	500
Start	99.0	94.0	89.0	68.0	49.0	48.0	67.0	50.0
Middle	75.0	64.0	62.0	70.0	75.0	69.0	71.0	83.0
End	94.0	83.0	94.0	95.0	98.0	96.0	97.0	97.0

Figure 11: Complete distractor boundary for Gemma4-26B-A4B: all eight counts and three target positions. The numeric panel reports accuracy (%). There are 100 examples per condition in a 1,000-record context. Horizontal positions are discrete tested counts.



Target / D	2	5	10	20	50	100	200	500
Start	86.0	86.0	89.0	92.0	74.0	56.0	52.0	16.0
Middle	66.0	31.0	29.0	20.0	24.0	41.0	61.0	72.0
End	92.0	78.0	53.0	47.0	32.0	30.0	40.0	29.0

Figure 12: Complete distractor boundary for Ministral-3-8B: all eight counts and three target positions. The numeric panel reports accuracy (%). There are 100 examples per condition in a 1,000-record context. Horizontal positions are discrete tested counts.

D.2 Complete Single-Position Salience-Marker Results

Purpose and design. This experiment asks whether the same target binding is retrieved differently after a textual intervention, and whether that change depends on the marker's location. Each model crosses four families with three marker locations and three target positions: 36 marked conditions. Qwen3-8B uses 450 records and 225 distractors; the other models use 1,000 records and 10 distractors. The target indices are 0/225/449 and 0/500/999, respectively. Start, Middle, and End locations are varied independently of these target choices.

Table 1 reports all 45 importance conditions in the main text; Tables 7, 8, and 9 report the remaining 135 conditions. No results are averaged across target positions. Parentheses show the change from the corresponding no-marker baseline; both accuracy and gain are bold only for a positive gain.

Secondary patterns. Strong middle-marker gains extend across all four families in the three Qwen models and Ministral, while start and end placements produce smaller or more variable changes. Gemma4 shows the opposite middle-marker effect across families. The non-valenced importance cue is among the strongest conditions in some models, while affectively framed cues are larger in others. Thus, the shared placement pattern further supports a general salience interpretation, with the direction and magnitude depending on model and family.

Table 7: Threat-marker accuracy (%) and change from no marker (pp). A positive gain makes both the accuracy and its parenthesis bold.

Model	Target position	No marker (%)	Threat marker location					
			Start		Middle		End	
Qwen3-8B	Start	54.0	**65.5**	**(+11.5)**	50.0	(−4.0)	**54.5**	**(+0.5)**
	Middle	63.0	**64.5**	**(+1.5)**	**78.5**	**(+15.5)**	**71.0**	**(+8.0)**
	End	58.0	**59.0**	**(+1.0)**	**100.0**	**(+42.0)**	**71.0**	**(+13.0)**
Qwen3.5-9B	Start	95.0	**97.0**	**(+2.0)**	**100.0**	**(+5.0)**	**99.0**	**(+4.0)**
	Middle	56.0	55.0	(−1.0)	**81.0**	**(+25.0)**	**64.0**	**(+8.0)**
	End	66.0	61.0	(−5.0)	**100.0**	**(+34.0)**	**67.0**	**(+1.0)**
Qwen3.6-35B-A3B	Start	100.0	100.0	(0.0)	100.0	(0.0)	100.0	(0.0)
	Middle	65.0	63.0	(−2.0)	**96.0**	**(+31.0)**	**66.0**	**(+1.0)**
	End	40.0	39.0	(−1.0)	**100.0**	**(+60.0)**	**51.0**	**(+11.0)**
Gemma4-26B-A4B	Start	89.0	81.0	(−8.0)	70.0	(−19.0)	69.0	(−20.0)
	Middle	62.0	**66.0**	**(+4.0)**	32.0	(−30.0)	**75.0**	**(+13.0)**
	End	94.0	92.0	(−2.0)	65.0	(−29.0)	**95.0**	**(+1.0)**
Ministral-3-8B	Start	89.0	87.0	(−2.0)	**92.0**	**(+3.0)**	84.0	(−5.0)
	Middle	29.0	29.0	(0.0)	**84.0**	**(+55.0)**	**33.0**	**(+4.0)**
	End	53.0	**62.0**	**(+9.0)**	**100.0**	**(+47.0)**	**84.0**	**(+31.0)**

Table 8: Reward-marker accuracy (%) and change from no marker (pp). Positive-gain conditions are bold, including the gain; baseline and nonpositive changes are not bold.

Model	Target position	No marker (%)	Reward marker location					
			Start		Middle		End	
Qwen3-8B	Start	54.0	**61.0**	**(+7.0)**	51.0	(−3.0)	**55.5**	**(+1.5)**
	Middle	63.0	**63.5**	**(+0.5)**	**71.0**	**(+8.0)**	**66.0**	**(+3.0)**
	End	58.0	**62.5**	**(+4.5)**	**100.0**	**(+42.0)**	58.0	(0.0)
Qwen3.5-9B	Start	95.0	**97.0**	**(+2.0)**	**99.0**	**(+4.0)**	**100.0**	**(+5.0)**
	Middle	56.0	**57.0**	**(+1.0)**	**81.0**	**(+25.0)**	56.0	(0.0)
	End	66.0	61.0	(−5.0)	**100.0**	**(+34.0)**	64.0	(−2.0)
Qwen3.6-35B-A3B	Start	100.0	100.0	(0.0)	100.0	(0.0)	100.0	(0.0)
	Middle	65.0	62.0	(−3.0)	**98.0**	**(+33.0)**	61.0	(−4.0)
	End	40.0	35.0	(−5.0)	**100.0**	**(+60.0)**	**51.0**	**(+11.0)**
Gemma4-26B-A4B	Start	89.0	86.0	(−3.0)	60.0	(−29.0)	72.0	(−17.0)
	Middle	62.0	**73.0**	**(+11.0)**	41.0	(−21.0)	**67.0**	**(+5.0)**
	End	94.0	**95.0**	**(+1.0)**	53.0	(−41.0)	**95.0**	**(+1.0)**
Ministral-3-8B	Start	89.0	81.0	(−8.0)	87.0	(−2.0)	86.0	(−3.0)
	Middle	29.0	29.0	(0.0)	**70.0**	**(+41.0)**	28.0	(−1.0)
	End	53.0	**61.0**	**(+8.0)**	**100.0**	**(+47.0)**	**70.0**	**(+17.0)**

Table 9: Empathy-marker accuracy (%) and change from no marker (pp). Positive-gain conditions are bold, including the gain. Values are not averaged across targets.

Model	Target position	No marker (%)	Empathy marker location					
			Start		Middle		End	
Qwen3-8B	Start	54.0	**59.5**	**(+5.5)**	50.5	(−3.5)	**66.0**	**(+12.0)**
	Middle	63.0	**70.0**	**(+7.0)**	**94.0**	**(+31.0)**	**65.5**	**(+2.5)**
	End	58.0	**65.5**	**(+7.5)**	**100.0**	**(+42.0)**	**59.5**	**(+1.5)**
Qwen3.5-9B	Start	95.0	**97.0**	**(+2.0)**	**99.0**	**(+4.0)**	**98.0**	**(+3.0)**
	Middle	56.0	56.0	(0.0)	**81.0**	**(+25.0)**	**59.0**	**(+3.0)**
	End	66.0	**67.0**	**(+1.0)**	**100.0**	**(+34.0)**	60.0	(−6.0)
Qwen3.6-35B-A3B	Start	100.0	100.0	(0.0)	100.0	(0.0)	100.0	(0.0)
	Middle	65.0	64.0	(−1.0)	**99.0**	**(+34.0)**	57.0	(−8.0)
	End	40.0	**43.0**	**(+3.0)**	**100.0**	**(+60.0)**	39.0	(−1.0)
Gemma4-26B-A4B	Start	89.0	83.0	(−6.0)	72.0	(−17.0)	82.0	(−7.0)
	Middle	62.0	**64.0**	**(+2.0)**	46.0	(−16.0)	**72.0**	**(+10.0)**
	End	94.0	94.0	(0.0)	74.0	(−20.0)	93.0	(−1.0)
Ministral-3-8B	Start	89.0	84.0	(−5.0)	86.0	(−3.0)	**90.0**	**(+1.0)**
	Middle	29.0	**31.0**	**(+2.0)**	**87.0**	**(+58.0)**	27.0	(−2.0)
	End	53.0	**62.0**	**(+9.0)**	**100.0**	**(+47.0)**	**70.0**	**(+17.0)**

D.3 Salience-Marker Position Combinations

Purpose and protocol. A strong single-location effect does not establish that repeated cues add further benefit. We therefore repeat the same family at Start+Middle, Start+End, Middle+End, or All positions, crossing the four combinations with four families and three target positions. This gives 48 combination conditions per model and 240 in total. The underlying records, distractor settings, query, sample counts, and decoding match the single-position experiment. Different marker families are not mixed within an example.

Two reference points. For model m, target position t, family e, and a set of marker locations C, let $A_{m,t,e}(C)$ denote accuracy in percent. We distinguish improvement over the unmarked condition from improvement over the best single placement:

$$\Delta_{\text{base}}(C) = A_{m,t,e}(C) - A_{m,t}(\emptyset), \tag{1}$$

$$G_{m,t,e}(C) = A_{m,t,e}(C) - \max_{p \in \{\text{Start},\text{Middle},\text{End}\}} A_{m,t,e}(\{p\}). \tag{2}$$

The maximum is computed within the same model, target, and family. We compute G for each family before averaging across families. Positive G indicates that the combination outperforms the strongest single placement for the same family.

Conditional additional gains. For middle targets (Table 10), Start+Middle and Middle+End exceed the best single placement by 11.4 and 7.9 pp on average in Qwen3-8B; Middle+End adds 9.3 pp in Ministral. The strongest average gains in Qwen3.5 and Qwen3.6 are smaller, 2.3 and 1.5 pp. Start+End is below the best single placement on average in all five models. Gemma4's Start+End nevertheless exceeds its no-marker baseline by 5.5 pp, while combinations involving Middle reduce baseline accuracy. These observations make the comparator essential: a combination can outperform no marker and still underperform a single marker.

Complete conditions and interpretation. The complete results are reported for Qwen3-8B (Table 11), Qwen3.5-9B (Table 12), Qwen3.6-35B-A3B (Table 13), Gemma4-26B-A4B (Table 14), and Ministral-3-8B (Table 15). These tables also show how the combination effect depends on target position. Final targets often reach ceiling accuracy with a single marker, leaving little room for further gains. At beginning targets, adding sites can reduce accuracy, whereas middle targets show the clearest differences between configurations. Bold identifies the highest accuracy in each row, including the no-marker and best-single references and ties.

Table 10: Middle-target combination effects averaged across the four families. Each entry is Δ_{base}/G in pp. S/M/E denote Start/Middle/End; G is computed within each family before averaging.

Model	S+M	S+E	M+E	All
Qwen3-8B	$+34.1\,/\,+11.4$	$+7.6\,/\,-15.1$	$+30.6\,/\,+7.9$	$+30.6\,/\,+7.9$
Qwen3.5-9B	$+25.5\,/\,-1.3$	$+14.5\,/\,-12.3$	$+28.8\,/\,+2.0$	$+29.0\,/\,+2.3$
Qwen3.6-35B-A3B	$+32.3\,/\,0.0$	$+2.5\,/\,-29.8$	$+33.8\,/\,+1.5$	$+32.0\,/\,-0.3$
Gemma4-26B-A4B	$-18.8\,/\,-29.0$	$+5.5\,/\,-4.8$	$-14.5\,/\,-24.8$	$-13.8\,/\,-24.0$
Ministral-3-8B	$+49.0\,/\,-0.8$	$+1.8\,/\,-48.0$	$+59.0\,/\,+9.3$	$+55.8\,/\,+6.0$

Table 11: Complete position-combination results for Qwen3-8B: all 48 conditions. S/M/E denote Start/Middle/End; values are accuracy (%). Best single is computed within the same target and family. Bold marks the row maximum, including references and ties.

Target	Family	No marker	Best single	S+M	S+E	M+E	All
Start	Threat	54.0	**65.5**	58.5	53.0	54.5	49.5
	Reward	54.0	61.0	52.0	**61.5**	53.0	57.5
	Empathy	54.0	**66.0**	57.0	55.5	63.5	56.5
	Importance	54.0	**66.5**	63.5	41.0	48.0	48.5
Middle	Threat	63.0	78.5	**97.0**	71.0	92.5	92.0
	Reward	63.0	71.0	**95.0**	71.5	84.0	85.5
	Empathy	63.0	94.0	**100.0**	65.0	98.0	97.5
	Importance	63.0	99.5	96.5	75.0	**100.0**	99.5
End	Threat	58.0	**100.0**	**100.0**	67.0	**100.0**	**100.0**
	Reward	58.0	**100.0**	99.5	70.0	**100.0**	**100.0**
	Empathy	58.0	**100.0**	**100.0**	69.0	**100.0**	**100.0**
	Importance	58.0	**100.0**	**100.0**	77.0	**100.0**	99.5

Table 12: Complete position-combination results for Qwen3.5-9B: all 48 conditions. S/M/E denote Start/Middle/End; values are accuracy (%). Best single is computed within the same target and family. Bold marks the row maximum, including references and ties.

Target	Family	No marker	Best single	S+M	S+E	M+E	All
Start	Threat	95.0	**100.0**	**100.0**	99.0	**100.0**	**100.0**
	Reward	95.0	**100.0**	99.0	**100.0**	98.0	**100.0**
	Empathy	95.0	99.0	99.0	**100.0**	99.0	**100.0**
	Importance	95.0	99.0	99.0	**100.0**	98.0	**100.0**
Middle	Threat	56.0	81.0	82.0	79.0	**93.0**	**93.0**
	Reward	56.0	**81.0**	**81.0**	64.0	74.0	80.0
	Empathy	56.0	81.0	79.0	71.0	**86.0**	78.0
	Importance	56.0	88.0	84.0	68.0	86.0	**89.0**
End	Threat	66.0	**100.0**	99.0	94.0	**100.0**	**100.0**
	Reward	66.0	**100.0**	99.0	70.0	**100.0**	99.0
	Empathy	66.0	**100.0**	**100.0**	86.0	**100.0**	**100.0**
	Importance	66.0	**100.0**	**100.0**	73.0	**100.0**	**100.0**

Table 13: Complete position-combination results for Qwen3.6-35B-A3B: all 48 conditions. S/M/E denote Start/Middle/End; values are accuracy (%). Best single is computed within the same target and family. Bold marks the row maximum, including references and ties.

Target	Family	No marker	Best single	S+M	S+E	M+E	All
Start	Threat	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	Reward	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	Empathy	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	Importance	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Middle	Threat	65.0	96.0	98.0	66.0	**99.0**	96.0
	Reward	65.0	98.0	97.0	64.0	**99.0**	98.0
	Empathy	65.0	99.0	99.0	72.0	**100.0**	99.0
	Importance	65.0	96.0	95.0	68.0	**97.0**	95.0
End	Threat	40.0	**100.0**	**100.0**	48.0	**100.0**	**100.0**
	Reward	40.0	**100.0**	**100.0**	46.0	**100.0**	**100.0**
	Empathy	40.0	**100.0**	**100.0**	38.0	**100.0**	**100.0**
	Importance	40.0	**100.0**	**100.0**	46.0	**100.0**	**100.0**

Table 14: Complete position-combination results for Gemma4-26B-A4B: all 48 conditions. S/M/E denote Start/Middle/End; values are accuracy (%). Best single is computed within the same target and family. Bold marks the row maximum, including references and ties.

Target	Family	No marker	Best single	S+M	S+E	M+E	All
Start	Threat	**89.0**	81.0	52.0	66.0	54.0	55.0
	Reward	**89.0**	86.0	65.0	76.0	50.0	58.0
	Empathy	**89.0**	83.0	73.0	79.0	71.0	75.0
	Importance	**89.0**	85.0	75.0	73.0	71.0	68.0
Middle	Threat	62.0	**75.0**	43.0	67.0	47.0	51.0
	Reward	62.0	**73.0**	41.0	70.0	56.0	55.0
	Empathy	62.0	**72.0**	46.0	68.0	45.0	48.0
	Importance	62.0	**69.0**	43.0	65.0	42.0	39.0
End	Threat	94.0	95.0	64.0	**98.0**	72.0	74.0
	Reward	94.0	95.0	67.0	**97.0**	72.0	77.0
	Empathy	**94.0**	**94.0**	86.0	88.0	81.0	88.0
	Importance	94.0	93.0	57.0	**95.0**	72.0	69.0

Table 15: Complete position-combination results for Ministral-3-8B: all 48 conditions. S/M/E denote Start/Middle/End; values are accuracy (%). Best single is computed within the same target and family. Bold marks the row maximum, including references and ties.

Target	Family	No marker	Best single	S+M	S+E	M+E	All
Start	Threat	89.0	**92.0**	88.0	78.0	91.0	83.0
	Reward	**89.0**	87.0	84.0	82.0	88.0	82.0
	Empathy	89.0	**90.0**	79.0	86.0	88.0	78.0
	Importance	**89.0**	87.0	82.0	74.0	77.0	67.0
Middle	Threat	29.0	84.0	86.0	34.0	**94.0**	88.0
	Reward	29.0	70.0	65.0	26.0	**79.0**	75.0
	Empathy	29.0	87.0	86.0	34.0	**88.0**	**88.0**
	Importance	29.0	74.0	75.0	29.0	**91.0**	88.0
End	Threat	53.0	**100.0**	**100.0**	92.0	**100.0**	**100.0**
	Reward	53.0	**100.0**	**100.0**	69.0	**100.0**	**100.0**
	Empathy	53.0	**100.0**	**100.0**	70.0	**100.0**	**100.0**
	Importance	53.0	**100.0**	**100.0**	70.0	**100.0**	**100.0**

D.4 Retrieval of Unmarked Targets: Complete Position Sweeps

Purpose and protocol. This test asks whether enhancing access around a marker systematically impairs access elsewhere. Marker location is fixed within a sweep while the queried record changes. Qwen3-8B uses 450 records, 200 examples per condition, and marker anchors 0, 225, and 450. Anchor 450 is the insertion site after the final record, not a target index. The other four models use 1,000 records, 100 examples per condition, and anchors 0, 500, and 999. Qwen3-8B is evaluated with all four families; the other models are evaluated with importance and empathy.

Display and comparison. For each fixed marker location, the first row gives baseline accuracy and subsequent rows give the family-specific change in percentage points. The horizontal labels are target offsets from the marker anchor, with one column per tested position. A boxed zero-offset column identifies the aligned target. Qwen3-8B's end anchor lies after the final record and therefore has no aligned target.

Qwen3-8B. The aligned middle target exhibits the strongest local improvements, while other target positions show smaller changes of both signs. The four-family comparison shows that the spatial pattern extends beyond importance and empathy (Figure 13).

Δ (pp) -60 0 +60

Marker at start (anchor = 0)

	+5	+10	+20	+50	+100	+150	+200	+250	+300
Baseline (%)	44	68	86	96.5	91.5	75	69.5	56.5	67
Importance Δ	+0.5	-2.5	-0.5	0	-1	-2.5	-3	-2	-3
Empathy Δ	+1.5	-1	+1	-0.5	-0.5	-4.5	-7.5	-5.5	-5
Threat Δ	-1	-2.5	-0.5	-0.5	+1.5	-3.5	-0.5	-1.5	-4.5
Reward Δ	-1	-6	-2	+1.5	+0.5	-1.5	-3	+0.5	-4

Marker at middle (anchor = 225)

	-200	-150	-100	-50	-20	-10	-5	**0**	+5	+10	+20	+50	+100	+150	+200
Baseline (%)	94.5	90	81	78	66	62.5	68.5	68	64	65.5	58	68.5	72	85.5	69.5
Importance Δ	-3	0	0	+2.5	+4	+6.5	+7	+32	+10.5	+9	+4	+0.5	+2	0	+5
Empathy Δ	-1	+2.5	+3	+0.5	+5.5	+7.5	+7	+27	+5.5	+4	+2.5	+2	-0.5	+0.5	+6.5
Threat Δ	0	+3.5	+4.5	+3.5	+6.5	+8	+6.5	+16	+4	+3	+3.5	+2	+2	0	+8.5
Reward Δ	0	+3	+3	+1.5	+7.5	+7.5	+7	+4	-3	-2.5	+0.5	+2.5	+1.5	0	+8.5

Marker at end (anchor = 450)

	-450	-400	-350	-300	-250	-200	-150	-100	-50
Baseline (%)	53.5	96	92	74.5	69.5	56.5	67	70.5	79
Importance Δ	-1	+0.5	-0.5	-0.5	-0.5	0	+0.5	+0.5	0
Empathy Δ	0	+0.5	-0.5	-0.5	-0.5	-0.5	-0.5	+0.5	+1.5
Threat Δ	0	+0.5	-0.5	-0.5	-0.5	-0.5	-0.5	+0.5	+1.5
Reward Δ	-1	+0.5	-0.5	-0.5	-0.5	0	+0.5	+0.5	0

Target position relative to marker (equally spaced tested positions)

Figure 13: Complete fixed-marker retrieval sweeps for Qwen3-8B. Baseline rows show accuracy (%); other rows show marker-induced changes (pp). The boxed column is aligned where such a target exists.

Exact target positions. For marker anchor 0: 5, 10, 20, 50, 100, 150, 200, 250, 300. For marker anchor 225: 25, 75, 125, 175, 205, 215, 220, 225, 230, 235, 245, 275, 325, 375, 425. For marker anchor 450: 400, 350, 300, 250, 200, 150, 100, 50, 0. The lists are record indices; the heatmaps label their offsets from the marker.

Qwen3.5-9B. The middle-aligned target improves under both families, but several neighboring or nonaligned targets decline. The sign therefore depends on the queried position as well as the marker anchor (Figure 14).

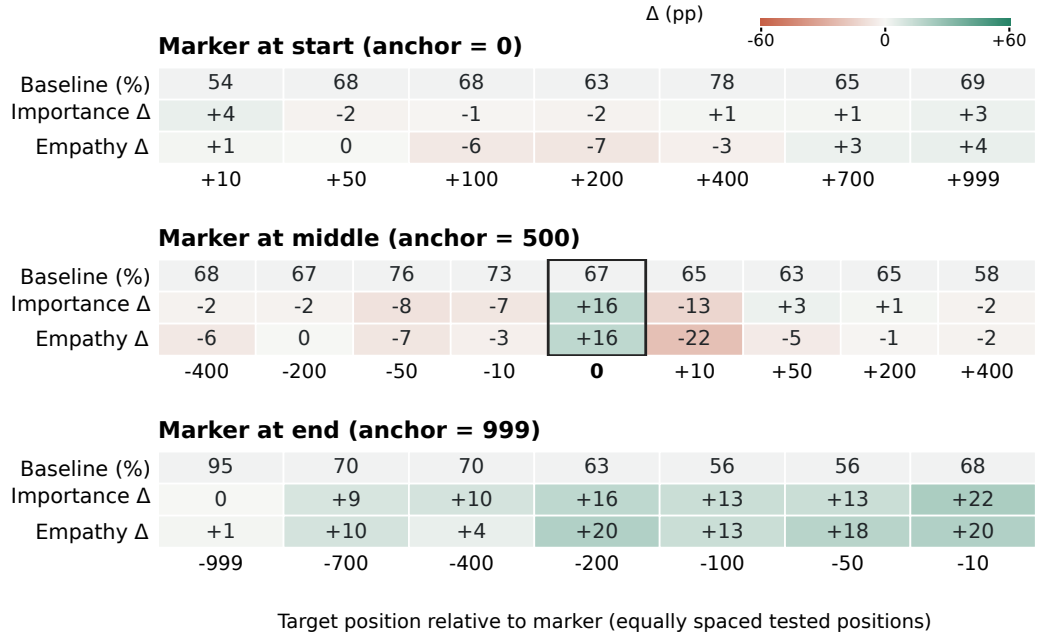


Figure 14: Complete fixed-marker retrieval sweeps for Qwen3.5-9B. Baseline rows show accuracy (%); other rows show marker-induced changes (pp). The boxed column is aligned where such a target exists.

Exact target positions. For marker anchor 0: 10, 50, 100, 200, 400, 700, 999. For marker anchor 500: 100, 300, 450, 490, 500, 510, 550, 700, 900. For marker anchor 999: 0, 299, 599, 799, 899, 949, 989. The lists are record indices; the heatmaps label their offsets from the marker.

Qwen3.6-35B-A3B. The middle-aligned condition produces a large retrieval gain, while changes at other targets vary with their position relative to the marker (Figure 15).

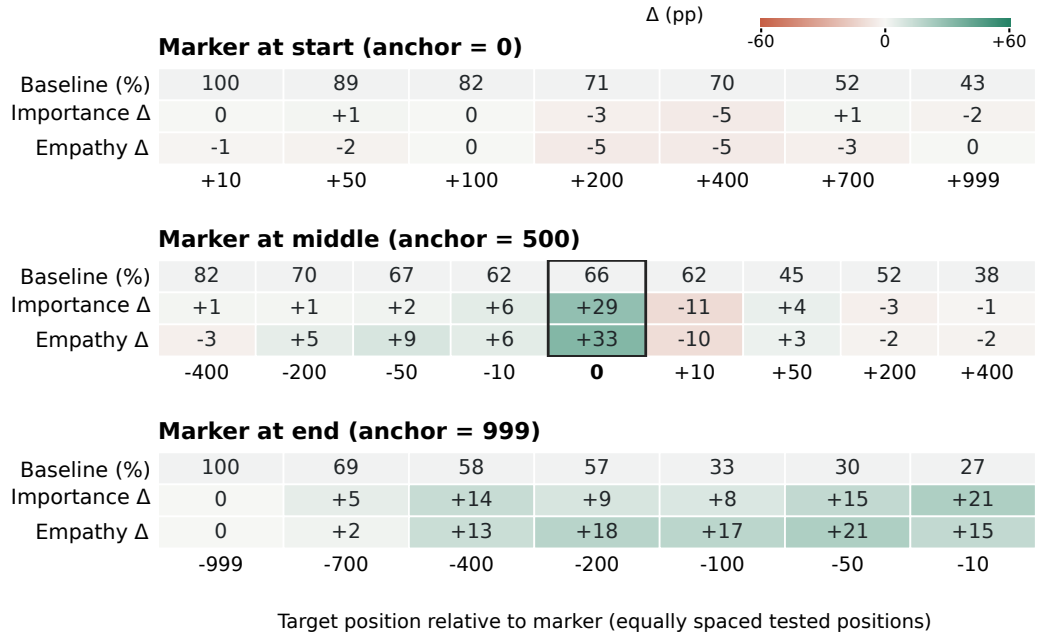


Figure 15: Complete fixed-marker retrieval sweeps for Qwen3.6-35B-A3B. Baseline rows show accuracy (%); other rows show marker-induced changes (pp). The boxed column is aligned where such a target exists.

Exact target positions. For marker anchor 0: 10, 50, 100, 200, 400, 700, 999. For marker anchor 500: 100, 300, 450, 490, 500, 510, 550, 700, 900. For marker anchor 999: 0, 299, 599, 799, 899, 949, 989. The lists are record indices; the heatmaps label their offsets from the marker.

Gemma4-26B-A4B. Middle marking reduces retrieval at the aligned target, making this model an informative counter-pattern. The complete sweep also separates that local reversal from effects at other targets (Figure 16).

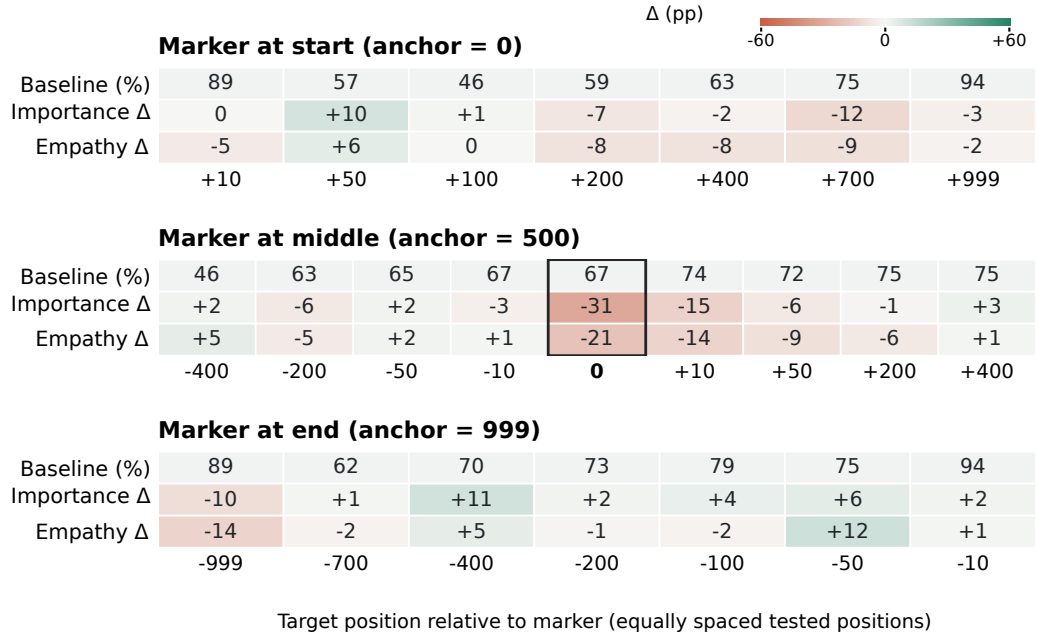


Figure 16: Complete fixed-marker retrieval sweeps for Gemma4-26B-A4B. Baseline rows show accuracy (%); other rows show marker-induced changes (pp). The boxed column is aligned where such a target exists.

Exact target positions. For marker anchor 0: 10, 50, 100, 200, 400, 700, 999. For marker anchor 500: 100, 300, 450, 490, 500, 510, 550, 700, 900. For marker anchor 999: 0, 299, 599, 799, 899, 949, 989. The lists are record indices; the heatmaps label their offsets from the marker.

Ministral-3-8B. Large gains occur around the middle marker and at several neighboring targets, alongside position-specific changes elsewhere. The effect extends to several nearby records while remaining position dependent (Figure 17).

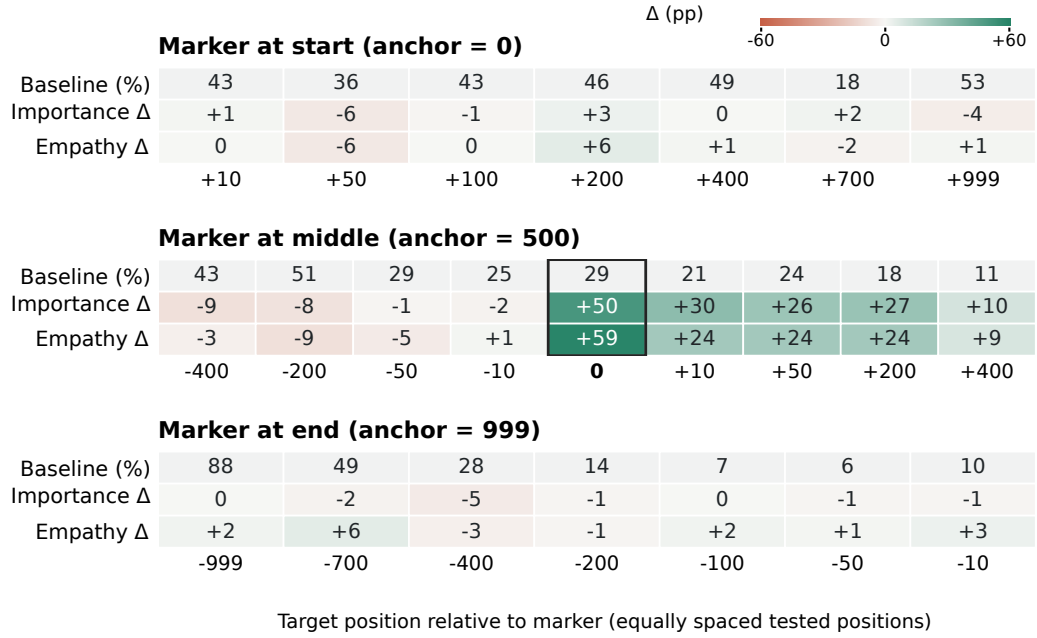


Figure 17: Complete fixed-marker retrieval sweeps for Ministral-3-8B. Baseline rows show accuracy (%); other rows show marker-induced changes (pp). The boxed column is aligned where such a target exists.

Exact target positions. For marker anchor 0: 10, 50, 100, 200, 400, 700, 999. For marker anchor 500: 100, 300, 450, 490, 500, 510, 550, 700, 900. For marker anchor 999: 0, 299, 599, 799, 899, 949, 989. The lists are record indices; the heatmaps label their offsets from the marker.

D.5 Memory Updating: Complete Family and Model Results

Purpose and protocol. A selective retrieval cue could still be harmful if it makes an old binding persist after a correction. To test this possibility, the same key receives $V_1 \rightarrow V_2$ in the single-update setting or $V_1 \rightarrow V_2 \rightarrow V_3$ in the multiple-update setting. The query always requests the latest value. We compare no marker, a marker on the old binding V_1, and a marker on the latest binding. The intermediate binding in the multiple-update condition is not the old-binding intervention target.

Qwen3-8B uses 450 records, target-position settings 0 and 225, and 200 examples per condition. Its single-update gaps are 50, 100, and 200 records, and multiple-update gaps are 50 and 100. The other models use 1,000 records, target-position settings 0 and 500, and 100 examples per condition; their single-update gaps are 100, 200, and 400, and multiple-update gaps are 100 and 200. Gaps are the record-gap settings in the experiment, not token distances. Each model includes all four families and all three marking conditions.

Reading the disaggregated results. Each model is shown with one panel per marker family. Within a panel, the Start/single, Start/multiple, Middle/single, and Middle/multiple settings form separate curves across update gaps. Every point measures latest-value accuracy. The accompanying table reports each condition in the order *no marker / old-binding marker / latest-binding marker*, in percent.

Retrieval under difficult update conditions. Each family-specific evaluation uses its paired no-marker control. Update interference can make the unmarked task particularly difficult: Qwen3.5 reaches 0% accuracy for a middle target after two updates at gap 200 in the empathy-paired condition, and Ministral reaches 0% for a middle target after one update at gap 100 in the importance-paired condition. Marking the latest binding raises these results to 98% and 68%, respectively. These cases show that a retrieval cue can restore access to updated information when the unmarked model fails throughout the evaluated sample.

Secondary conclusions. Qwen3-8B maintains high latest-value accuracy when the latest binding is marked, and Qwen3.6 remains near ceiling in many settings. Qwen3.5 and Ministral show larger gains in some difficult configurations. Gemma4 responds differently to single and multiple updates: latest-binding marking can reduce accuracy after one update while improving it after two. The update count and marker attachment jointly determine how the cue affects access to the current binding.

Qwen3-8B. Marking the latest binding maintains high accuracy across the four families. Old-binding marking is generally compatible with updating but has individual decreases, visible in the separate gap settings. Figure 18 shows the complete curves, and Table 16 reports the corresponding accuracies.

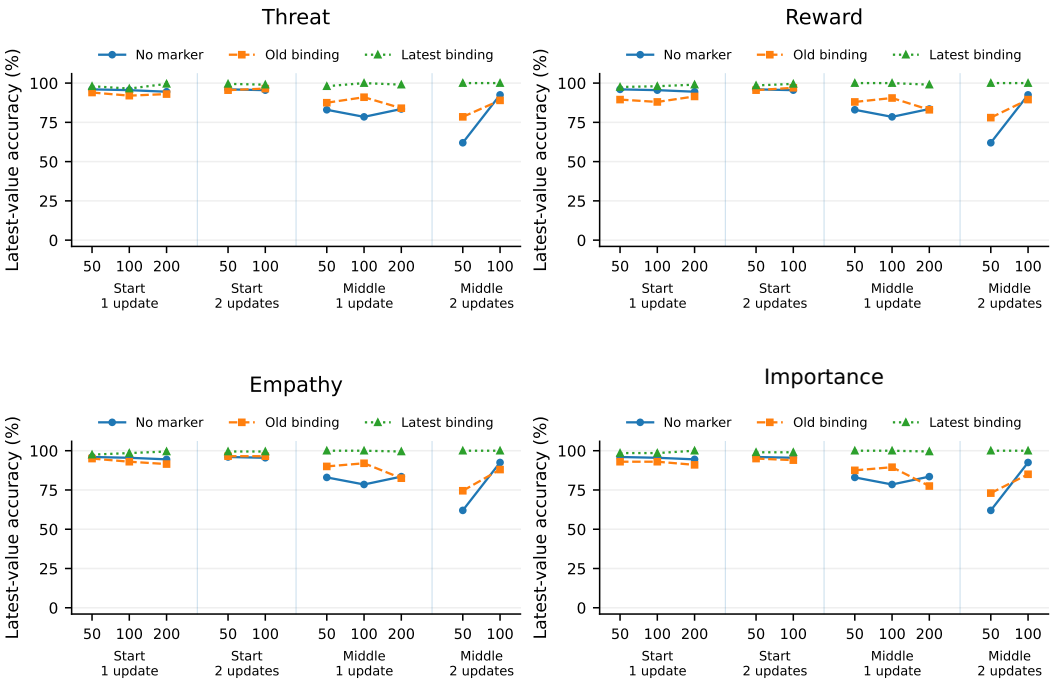


Figure 18: All four marker families and both update counts for Qwen3-8B. Each curve segment belongs to a separate target/update setting. Tick labels give record-gap settings. Styles distinguish no marker, old-binding marking, and latest-binding marking; every point evaluates the latest value.

Table 16: Exact update accuracies for Qwen3-8B (%). Each family cell lists no marker / old-binding marker / latest-binding marker. Each family is evaluated against its paired no-marker control.

Target	Updates	Gap	Threat	Reward	Empathy	Importance
Start	1	50	96.0 / 94.0 / 98.0	96.0 / 89.5 / 97.5	96.0 / 95.0 / 97.5	96.0 / 93.0 / 98.5
Start	1	100	95.5 / 92.0 / 96.5	95.5 / 88.0 / 98.0	95.5 / 93.0 / 98.5	95.5 / 93.0 / 98.5
Start	1	200	94.5 / 93.0 / 99.5	94.5 / 91.5 / 99.0	94.5 / 91.5 / 99.5	94.5 / 91.0 / 100.0
Start	2	50	96.0 / 95.5 / 99.5	96.0 / 95.5 / 98.5	96.0 / 96.5 / 99.5	96.0 / 95.0 / 99.0
Start	2	100	95.5 / 96.5 / 99.0	95.5 / 97.0 / 99.5	95.5 / 96.5 / 99.5	95.5 / 94.0 / 99.0
Middle	1	50	83.0 / 87.5 / 98.0	83.0 / 88.0 / 100.0	83.0 / 90.0 / 100.0	83.0 / 87.5 / 100.0
Middle	1	100	78.5 / 91.0 / 100.0	78.5 / 90.5 / 100.0	78.5 / 92.0 / 100.0	78.5 / 89.5 / 100.0
Middle	1	200	83.5 / 84.0 / 99.0	83.5 / 83.0 / 99.0	83.5 / 82.5 / 99.5	83.5 / 77.5 / 99.5
Middle	2	50	62.0 / 78.5 / 100.0	62.0 / 78.0 / 100.0	62.0 / 74.5 / 100.0	62.0 / 73.0 / 100.0
Middle	2	100	92.5 / 89.0 / 100.0	92.5 / 89.5 / 100.0	92.5 / 88.0 / 100.0	92.5 / 85.0 / 100.0

Qwen3.5-9B. The effects differ across update configurations. In the middle-target two-update setting, the no-marker condition can fail completely in some runs, including the 0% baseline observed for the empathy-paired condition at gap 200. Marker conditions can nevertheless recover high latest-value retrieval, illustrating that updating itself can become difficult without an additional retrieval cue. Figure 19 shows the complete curves, and Table 17 reports the corresponding accuracies.

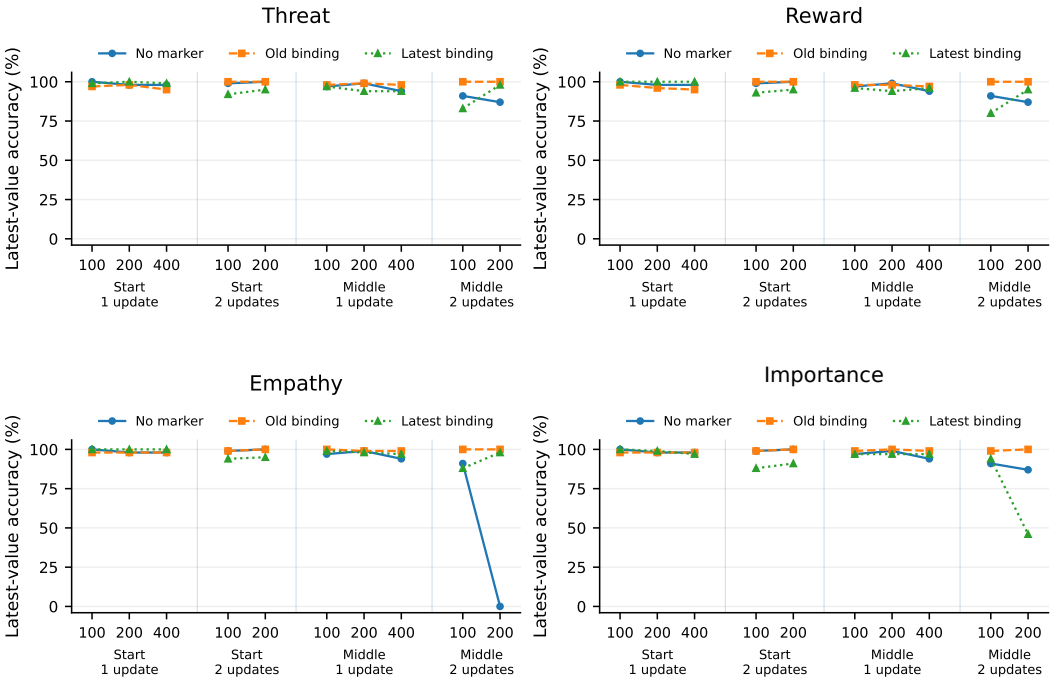


Figure 19: All four marker families and both update counts for Qwen3.5-9B. Each curve segment belongs to a separate target/update setting. Tick labels give record-gap settings. Styles distinguish no marker, old-binding marking, and latest-binding marking; every point evaluates the latest value.

Table 17: Exact update accuracies for Qwen3.5-9B (%). Each family cell lists no marker / old-binding marker / latest-binding marker. Each family is evaluated against its paired no-marker control.

Target	Updates	Gap	Threat	Reward	Empathy	Importance
Start	1	100	100.0 / 97.0 / 99.0	100.0 / 98.0 / 100.0	100.0 / 98.0 / 100.0	100.0 / 98.0 / 100.0
Start	1	200	98.0 / 98.0 / 100.0	98.0 / 96.0 / 100.0	98.0 / 98.0 / 100.0	98.0 / 98.0 / 99.0
Start	1	400	98.0 / 95.0 / 99.0	98.0 / 95.0 / 100.0	98.0 / 98.0 / 100.0	98.0 / 98.0 / 97.0
Start	2	100	99.0 / 100.0 / 92.0	99.0 / 100.0 / 93.0	99.0 / 99.0 / 94.0	99.0 / 99.0 / 88.0
Start	2	200	100.0 / 100.0 / 95.0	100.0 / 100.0 / 95.0	100.0 / 100.0 / 95.0	100.0 / 100.0 / 91.0
Middle	1	100	97.0 / 98.0 / 97.0	97.0 / 98.0 / 96.0	97.0 / 100.0 / 99.0	97.0 / 99.0 / 97.0
Middle	1	200	99.0 / 99.0 / 94.0	99.0 / 98.0 / 94.0	99.0 / 99.0 / 98.0	99.0 / 100.0 / 97.0
Middle	1	400	94.0 / 98.0 / 94.0	94.0 / 97.0 / 96.0	94.0 / 99.0 / 97.0	94.0 / 99.0 / 97.0
Middle	2	100	91.0 / 100.0 / 83.0	91.0 / 100.0 / 80.0	91.0 / 100.0 / 88.0	91.0 / 99.0 / 94.0
Middle	2	200	87.0 / 100.0 / 98.0	87.0 / 100.0 / 95.0	0.0 / 100.0 / 98.0	87.0 / 100.0 / 46.0

Qwen3.6-35B-A3B. Most conditions are near ceiling, leaving little room for improvement. The largest differences occur in the middle-target single-update conditions. Figure 20 shows the complete curves, and Table 18 reports the corresponding accuracies.

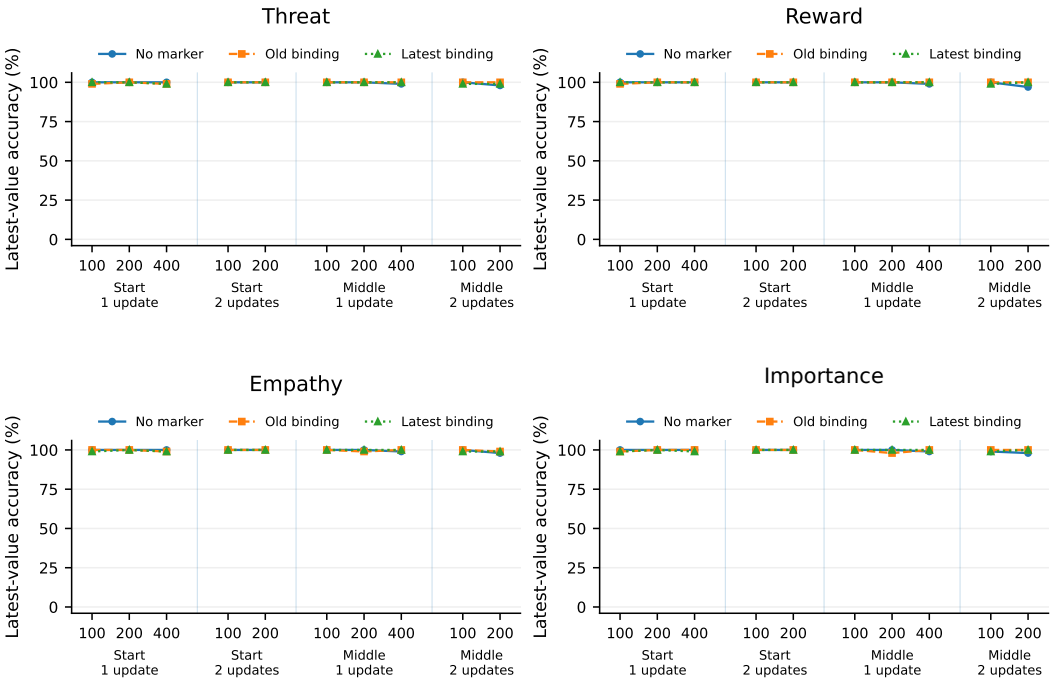


Figure 20: All four marker families and both update counts for Qwen3.6-35B-A3B. Each curve segment belongs to a separate target/update setting. Tick labels give record-gap settings. Styles distinguish no marker, old-binding marking, and latest-binding marking; every point evaluates the latest value.

Table 18: Exact update accuracies for Qwen3.6-35B-A3B (%). Each family cell lists no marker / old-binding marker / latest-binding marker. Each family is evaluated against its paired no-marker control.

Target	Updates	Gap	Threat	Reward	Empathy	Importance
Start	1	100	100.0 / 99.0 / 100.0	100.0 / 99.0 / 100.0	100.0 / 100.0 / 99.0	100.0 / 99.0 / 99.0
Start	1	200	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0
Start	1	400	100.0 / 99.0 / 99.0	100.0 / 100.0 / 100.0	100.0 / 99.0 / 99.0	100.0 / 100.0 / 99.0
Start	2	100	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0
Start	2	200	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0
Middle	1	100	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0
Middle	1	200	100.0 / 100.0 / 100.0	100.0 / 100.0 / 100.0	100.0 / 99.0 / 100.0	100.0 / 98.0 / 100.0
Middle	1	400	99.0 / 100.0 / 100.0	99.0 / 100.0 / 100.0	99.0 / 100.0 / 100.0	99.0 / 100.0 / 100.0
Middle	2	100	100.0 / 100.0 / 99.0	100.0 / 100.0 / 99.0	100.0 / 100.0 / 99.0	99.0 / 100.0 / 99.0
Middle	2	200	98.0 / 100.0 / 99.0	97.0 / 100.0 / 100.0	98.0 / 99.0 / 99.0	98.0 / 100.0 / 100.0

Gemma4-26B-A4B. Latest-binding marking can reduce accuracy for single updates yet help for multiple updates. At gap 200 in the middle-target two-update empathy condition, marking the latest binding raises accuracy from 14% to 97%. Figure 21 shows the complete curves, and Table 19 reports the corresponding accuracies.

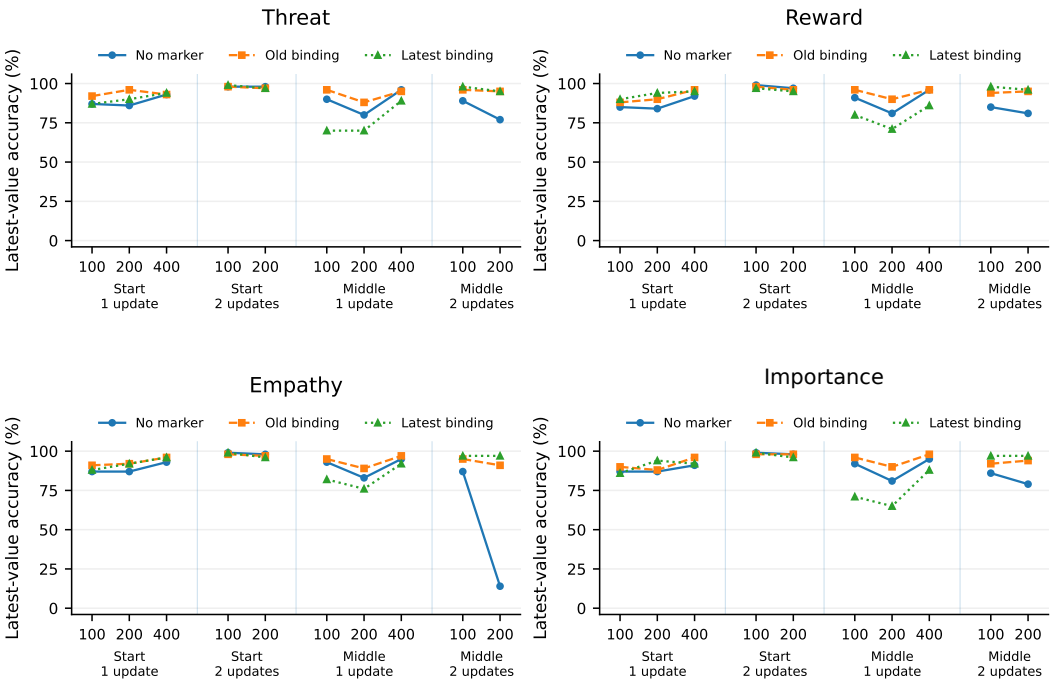


Figure 21: All four marker families and both update counts for Gemma4-26B-A4B. Each curve segment belongs to a separate target/update setting. Tick labels give record-gap settings. Styles distinguish no marker, old-binding marking, and latest-binding marking; every point evaluates the latest value.

Table 19: Exact update accuracies for Gemma4-26B-A4B (%). Each family cell lists no marker / old-binding marker / latest-binding marker. Each family is evaluated against its paired no-marker control.

Target	Updates	Gap	Threat	Reward	Empathy	Importance
Start	1	100	87.0 / 92.0 / 87.0	85.0 / 88.0 / 90.0	87.0 / 91.0 / 88.0	87.0 / 90.0 / 86.0
Start	1	200	86.0 / 96.0 / 90.0	84.0 / 90.0 / 94.0	87.0 / 92.0 / 92.0	87.0 / 88.0 / 94.0
Start	1	400	93.0 / 93.0 / 94.0	92.0 / 96.0 / 95.0	93.0 / 96.0 / 96.0	91.0 / 96.0 / 92.0
Start	2	100	98.0 / 98.0 / 99.0	99.0 / 98.0 / 97.0	99.0 / 98.0 / 99.0	99.0 / 98.0 / 99.0
Start	2	200	98.0 / 97.0 / 97.0	97.0 / 96.0 / 95.0	98.0 / 97.0 / 96.0	98.0 / 98.0 / 96.0
Middle	1	100	90.0 / 96.0 / 70.0	91.0 / 96.0 / 80.0	93.0 / 95.0 / 82.0	92.0 / 96.0 / 71.0
Middle	1	200	80.0 / 88.0 / 70.0	81.0 / 90.0 / 71.0	83.0 / 89.0 / 76.0	81.0 / 90.0 / 65.0
Middle	1	400	96.0 / 95.0 / 89.0	96.0 / 96.0 / 86.0	95.0 / 97.0 / 92.0	95.0 / 98.0 / 88.0
Middle	2	100	89.0 / 96.0 / 98.0	85.0 / 94.0 / 98.0	87.0 / 95.0 / 97.0	86.0 / 92.0 / 97.0
Middle	2	200	77.0 / 95.0 / 95.0	81.0 / 95.0 / 96.0	14.0 / 91.0 / 97.0	79.0 / 94.0 / 97.0

Ministral-3-8B. Latest-binding markers often substantially improve access to the updated value, particularly for middle targets. Some no-marker conditions are extremely difficult and can reach 0% accuracy, showing that the model may completely fail to retrieve the latest binding under update interference. Figure 22 shows the complete curves, and Table 20 reports the corresponding accuracies.

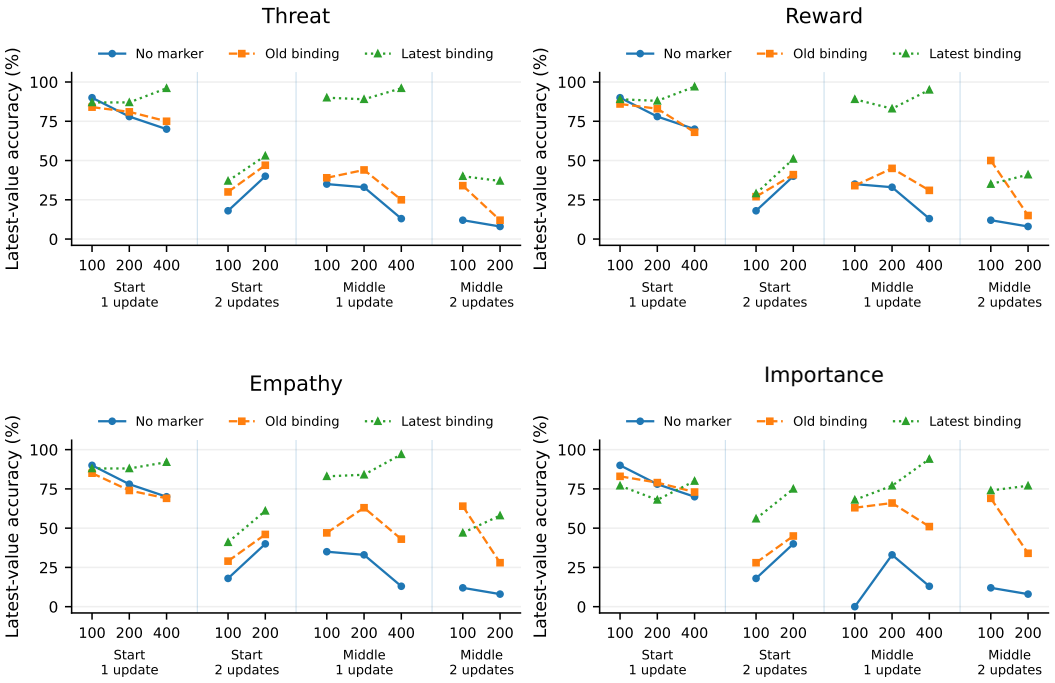


Figure 22: All four marker families and both update counts for Ministral-3-8B. Each curve segment belongs to a separate target/update setting. Tick labels give record-gap settings. Styles distinguish no marker, old-binding marking, and latest-binding marking; every point evaluates the latest value.

Table 20: Exact update accuracies for Ministral-3-8B (%). Each family cell lists no marker / old-binding marker / latest-binding marker. Each family is evaluated against its paired no-marker control.

Target	Updates	Gap	Threat	Reward	Empathy	Importance
Start	1	100	90.0 / 84.0 / 87.0	90.0 / 86.0 / 89.0	90.0 / 85.0 / 88.0	90.0 / 83.0 / 77.0
Start	1	200	78.0 / 81.0 / 87.0	78.0 / 83.0 / 88.0	78.0 / 74.0 / 88.0	78.0 / 79.0 / 68.0
Start	1	400	70.0 / 75.0 / 96.0	70.0 / 68.0 / 97.0	70.0 / 69.0 / 92.0	70.0 / 73.0 / 80.0
Start	2	100	18.0 / 30.0 / 37.0	18.0 / 27.0 / 29.0	18.0 / 29.0 / 41.0	18.0 / 28.0 / 56.0
Start	2	200	40.0 / 47.0 / 53.0	40.0 / 41.0 / 51.0	40.0 / 46.0 / 61.0	40.0 / 45.0 / 75.0
Middle	1	100	35.0 / 39.0 / 90.0	35.0 / 34.0 / 89.0	35.0 / 47.0 / 83.0	0.0 / 63.0 / 68.0
Middle	1	200	33.0 / 44.0 / 89.0	33.0 / 45.0 / 83.0	33.0 / 63.0 / 84.0	33.0 / 66.0 / 77.0
Middle	1	400	13.0 / 25.0 / 96.0	13.0 / 31.0 / 95.0	13.0 / 43.0 / 97.0	13.0 / 51.0 / 94.0
Middle	2	100	12.0 / 34.0 / 40.0	12.0 / 50.0 / 35.0	12.0 / 64.0 / 47.0	12.0 / 69.0 / 74.0
Middle	2	200	8.0 / 12.0 / 37.0	8.0 / 15.0 / 41.0	8.0 / 28.0 / 58.0	8.0 / 34.0 / 77.0

E Internal Analysis Details

E.1 Shared Configuration, Positions, and Measurement Scope

The main internal analyses compare a no-marker baseline with the importance string `[!!! IMPORTANT DATA !!!]` inserted at Middle, with a middle target. This non-valenced salience cue produces a strong behavioral effect and permits direct comparison of the internal diagnostics without requiring an affective interpretation. Table 21 specifies the main readout, attention, and steering cohorts. Each example contributes paired conditions; the low-precision-model backends are the proxies described in Appendix B.

Table 21: Main internal-analysis cohorts. N: records; D: similar-key distractors; n: paired examples. Target indices are zero based.

Model	N	D	Target	n
Qwen3-8B	450	225	225	200
Qwen3.5-9B	1,000	10	500	100
Qwen3.6-35B-A3B	1,000	10	500	100
Gemma4-26B-A4B	1,000	10	500	100
Ministral-3-8B	1,000	10	500	100

Retrieval-position alignment. Measurements are taken immediately before generation of the answer's first value token. Qwen3-8B can produce a natural-language prefix before the value. We retain each condition's natural prefix and locate its pre-value position separately, denoted t^*_B and t^*_E. The prefixes and absolute token indices can therefore differ even when the input memories are paired. The other models directly output raw values and use the prompt-final position. In a steering run, the baseline prefix is held fixed; the experiment intervenes on its continuation rather than replacing it with the marked-condition prefix.

Distinct diagnostic quantities. The analyses track successive aspects of retrieval: layer-wise readouts measure support for candidate first tokens, attention measures access to their source spans, cosine describes changes in feed-forward output direction, and steering measures the target-probability change caused by transferring a shared MLP direction. Behavioral accuracy evaluates the complete generated value.

E.2 Layer-wise Readouts and Fixed Candidate Controls

Logit Lens. At each layer, we apply the model's final normalization and language-model head to the residual state at the retrieval position (nostalgebraist, 2020). Writing the output projection as W_U and final normalization as Norm_f, the readout has the form

$$q^{\mathrm{LL}}_{i\ell} = \mathrm{softmax}(W_U \,\mathrm{Norm}_f(h_{i\ell}(t^*)))\,. \tag{3}$$

Readouts use each model's native output-head processing.

Jacobian Lens. Jacobian Lens accounts for downstream transformations by mapping the intermediate state through a layer-specific average Jacobian before applying the final normalization and head (Gurnee et al., 2026). The fitted matrices use an independent WikiText-103 corpus (Merity et al., 2016), rather than the retrieval evaluation examples. Schematically, the readout is $q^{\mathrm{JL}}_{i\ell} = \mathrm{softmax}(W_U \,\mathrm{Norm}_f(J_\ell h_{i\ell}(t^*)))$. Intermediate values are normalized readout scores, and the final LM head provides the actual output distribution. Both lenses are used to track the layer-wise separation of the fixed candidate groups.

Candidate construction. The target group contains the first token of the correct value. Mean D10 averages the readouts of ten similar-distractor value candidates; Mean N10 averages ten background-value candidates associated with unrelated keys. Fixed Early-NV10 contains ten nonvalue controls selected symmetrically from early-layer readouts of both conditions after excluding memory-value

first tokens and special tokens. Once selected, the groups are held fixed across layers and conditions. This avoids changing the comparison set across layers.

D10 samples ten candidates from the distractor values in the context. The corresponding attention category includes the full set of distractor sources. Group readouts are means over candidate first-token scores; candidate strings that share a first token contribute the score of that token.

Plots and uncertainty. At each layer, readouts are averaged over paired examples. The Qwen3-8B plots show bootstrap 95% confidence intervals. Logarithmic axes display both weak early-layer scores and their later separation. Solid lines denote baseline and dashed lines denote Middle importance; `P2` in the cross-model legend refers to the middle marker location.

E.3 Attention Aggregation and Layer Coverage

Source spans. At the retrieval position, we reconstruct each analyzed head's post-softmax attention row and assign context tokens to five source categories: target key, target value, the union of distractor keys, the union of distractor values, and the repeated key in the final query. The categories distinguish access to a binding's identity from access to its value. Marker tokens are excluded from the paired source-span difference because there is no corresponding marker text in the baseline. These categories exclude unrelated context tokens and are reported as separate source-span statistics.

Token-normalized aggregation. Let $a^c_{i\ell h}(t^*_c, s)$ denote attention from the retrieval position to source token s, for sample i, layer ℓ, head h, and condition $c \in \{B, E\}$, where B is baseline and E is the experimental importance-marker condition. For a span category S, we compute

$$A^c_{\ell,S} = \frac{1}{H_\ell} \sum_{h=1}^{H_\ell} \frac{\sum_i \sum_{s \in S^c_i} a^c_{i\ell h}(t^*_c, s)}{\sum_i |S^c_i|}, \qquad \Delta A_{\ell,S} = A^E_{\ell,S} - A^B_{\ell,S}. \tag{4}$$

Attention mass and source-token counts are pooled over examples before averaging heads. This pooled per-token statistic differs from an unweighted mean of per-example ratios when lengths differ, and controls for the size of a distractor span. Positive values indicate more attention per source token under importance. Table 22 lists the included layers for each model.

Table 22: Attention coverage. Full MLP-layer coverage is retained for the cosine and steering analyses even when the attention comparison includes only global/full-attention layers.

Model	Included attention layers
Qwen3-8B	L1–L36
Qwen3.5-9B	L4, L8, L12, L16, L20, L24, L28, L32
Qwen3.6-35B-A3B	L4, L8, . . ., L40
Gemma4-26B-A4B	Global layers L6, L12, L18, L24, L30
Ministral-3-8B	L1–L34

Gemma4's remaining attention layers use 1,024-token sliding windows; the long-range comparison uses its global layers. Annotations are in units of 10^{-3}, so a displayed $+4.94$ corresponds to an attention increase of 0.00494 per source token.

E.4 Paired MLP-Output Direction Similarity

Question and extraction point. The cosine analysis asks where a marker changes the direction of the MLP/FFN branch output on the same retrieval example. It precedes the stronger question of whether a transferable difference can influence an unmarked example. We read each layer's branch output at its condition-specific pre-value position, before residual addition. For Qwen3.6 this includes routed MoE and the gated shared expert; for Gemma4 it includes the combined dense-FFN/MoE output. The measurement is the branch output entering residual addition, excluding the pre-existing residual state.

Per-example statistic, then sample mean. For paired vectors $m_{i\ell}^B$ and $m_{i\ell}^E$, the plotted quantity is

$$c_{i\ell} = \frac{\langle m_{i\ell}^B, m_{i\ell}^E \rangle}{\|m_{i\ell}^B\|_2 \|m_{i\ell}^E\|_2}, \qquad \bar{c}_\ell = \frac{1}{n} \sum_{i=1}^{n} c_{i\ell}. \tag{5}$$

Each curve shows $\bar{c}_\ell$: cosine is computed within each example before averaging over examples. Values closer to one indicate more similar output directions. The layer tables additionally report sample medians, minima, and maxima.

Experimental configuration. Qwen3-8B uses 200 paired examples with 450 records, 225 distractors, and the target at index 225. The other four models use 100 pairs with 1,000 records, 10 distractors, and the target at index 500, using the backends in Appendix B. All conditions compare Middle importance with baseline. The cosine and steering results are summarized separately for each layer, covering 36, 32, 40, 30, and 34 layers, respectively.

E.5 Cross-fitted Shared MLP Directions

Direction construction and held-out transfer. Let $m_{i\ell}^B$ and $m_{i\ell}^E$ denote the pre-residual branch outputs at their respective pre-value positions. We first form $\delta_{i\ell} = m_{i\ell}^E - m_{i\ell}^B$. The samples are assigned to five folds, and the direction for held-out fold $\mathcal{F}_k$ is

$$v_{k\ell} = \frac{1}{|\mathcal{I} \setminus \mathcal{F}_k|} \sum_{i \notin \mathcal{F}_k} \delta_{i\ell}. \tag{6}$$

Each direction uses only the other 80% of examples: 160 in Qwen3-8B and 80 in each other model. The evaluated example's own marked-condition activation is excluded, testing shared transfer rather than same-example reconstruction. The direction is the raw mean difference, with unit additive coefficient; no per-example normalization is applied to the injected vector.

Intervention site and outcome. For $i \in \mathcal{F}_k$, one forward evaluation modifies one layer at one position:

$$\widetilde{m}_{i\ell}^B = m_{i\ell}^B + v_{k\ell}, \qquad \Delta P_{i\ell} = P_i(\mathrm{Target}_1 \mid \ell\text{-patched baseline}) - P_i(\mathrm{Target}_1 \mid \mathrm{baseline}). \tag{7}$$

The addition occurs before residual addition, only at t_B^* in layer ℓ, after which the remaining computation runs normally. Each intervention targets only this single layer and retrieval position. The reported mean pools held-out outcomes after cycling through the five folds. The full tables also give the median and the fraction of held-out examples with strictly positive ΔP.

Layer-wise intervention sweep. The intervention isolates a layer-specific, cross-sample change in the feed-forward output. Sweeping the intervention layer identifies where transferring this shared direction increases or decreases support for the target. The approach is related to activation-steering methods (Turner et al., 2023; Rimsky et al., 2024), with both measurement and injection at the pre-residual MLP/FFN output.

F Additional Internal Analysis Results

F.1 Cross-Model Layer-wise Readouts

Cross-model readouts compare Target, Mean D10, Mean N10, and Fixed Early-NV10 under baseline and Middle importance. Candidate construction follows Appendix E.2. The results are shown for Qwen3.5-9B (Figure 23), Qwen3.6-35B-A3B (Figure 24), Gemma4-26B-A4B (Figure 25), and Ministral-3-8B (Figure 26).

Shared transition, differing trajectories. Qwen3.5, Qwen3.6, and Ministral develop stronger target support in later layers, while Gemma4 follows a distinct intermediate trajectory. Across the two lenses, early nonvalue controls remain prominent before the candidate values separate. The timing of this transition varies across models, but in each case the largest value-specific separation develops toward the end of the layer stack.

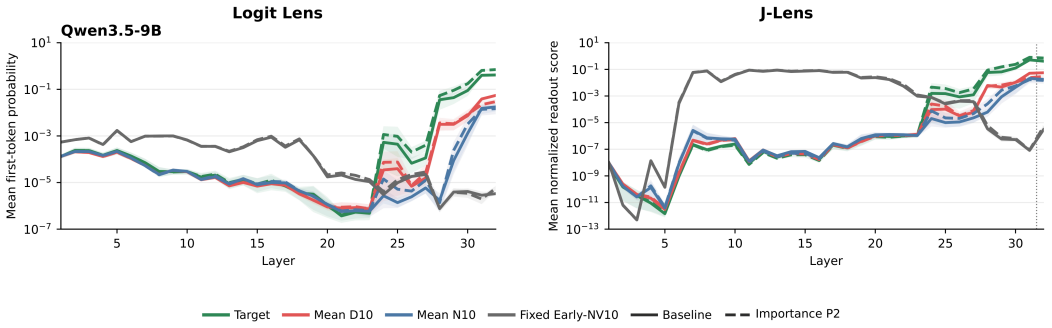


Figure 23: Cross-model Logit Lens (left) and Jacobian Lens (right): Qwen3.5-9B. Candidate groups are fixed across layers and conditions. Solid lines show baseline, dashed lines Middle importance. Jacobian Lens values are normalized readout scores.

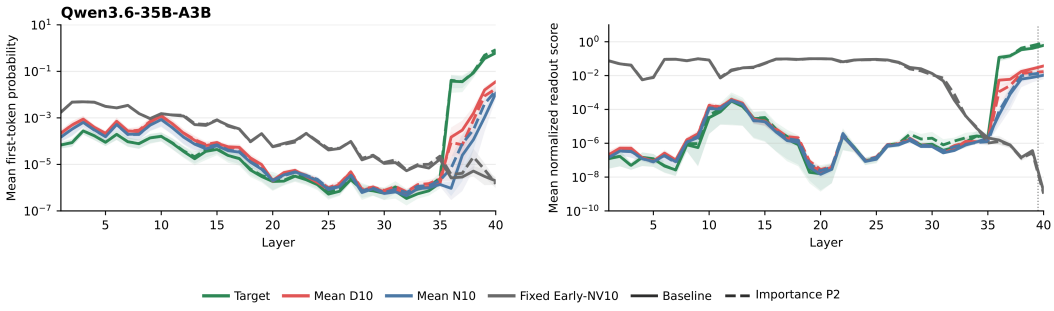


Figure 24: Cross-model Logit Lens (left) and Jacobian Lens (right): Qwen3.6-35B-A3B. Candidate groups are fixed across layers and conditions. Solid lines show baseline, dashed lines Middle importance. Jacobian Lens values are normalized readout scores.

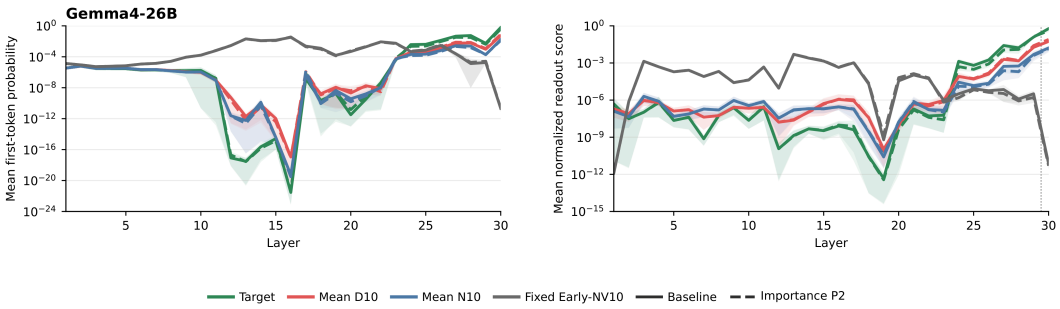


Figure 25: Cross-model Logit Lens (left) and Jacobian Lens (right): Gemma4-26B-A4B. Candidate groups are fixed across layers and conditions. Solid lines show baseline, dashed lines Middle importance. Jacobian Lens values are normalized readout scores.

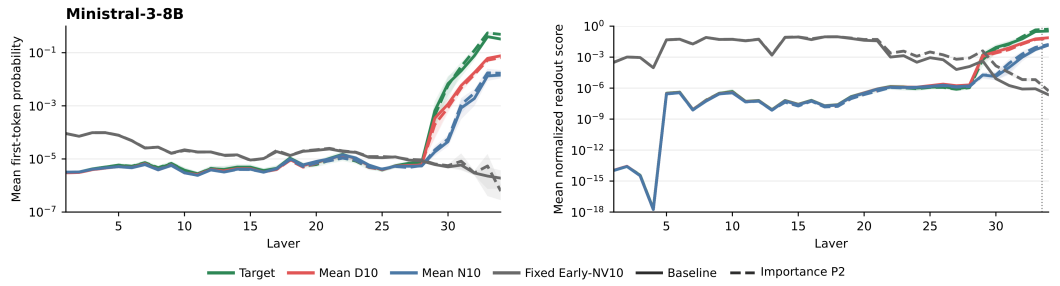


Figure 26: Cross-model Logit Lens (left) and Jacobian Lens (right): Ministral-3-8B. Candidate groups are fixed across layers and conditions. Solid lines show baseline, dashed lines Middle importance. Jacobian Lens values are normalized readout scores.

F.2 Cross-Model Attention Redistribution

The cross-model attention differences use Equation 4 and the layer coverage in Table 22. Qwen3-8B's target-value increase peaks at 4.94×10^{-3} in L25, while distractor-value attention declines over L10–L36 (Figure 27). The other Qwen models and Ministral show related target-oriented changes with local reversals (Figures 28, 29, and 31). Gemma4 instead reduces target-value attention by 2.81×10^{-3} at L24 and 1.01×10^{-3} at L30 while increasing distractor-value attention. The opposite target–distractor redistribution in Gemma4 (Figure 30) parallels its behavioral response to the same marker.

	Target key	Target value	D key	D value	Final query key		Target key	Target value	D key	D value	Final query key
L1	+1.65e-05	+5.02e-05	+0.00016	+0.000124	+0.0792	L19	+0.628	+0.299	-0.00103	-0.000648	-0.661
L2	+0.000327	+8.34e-05	+0.000237	+0.000154	+0.0187	L20	+0.258	+0.43	+0.00147	-0.00112	-0.762
L3	+0.00743	+0.00814	+0.000154	+4.47e-05	+0.0269	L21	+0.46	+2.91	+0.00167	-0.00823	-0.326
L4	+0.00875	+0.00144	-1.45e-05	+8.86e-06	-0.00532	L22	+0.642	+2.66	-0.00411	-0.0133	-1.23
L5	+0.012	+0.00112	+9.82e-05	+5.62e-05	-0.00786	L23	+0.795	+1.02	-0.000386	-0.00629	-2.21
L6	+0.0151	+0.00154	+0.000109	+0.000108	+0.0132	L24	+0.768	+4.18	-0.00811	-0.0295	-1.97
L7	+0.0245	+0.00422	-8.51e-06	+2.12e-05	-0.0193	L25	+1.02	+4.94	-0.00859	-0.0391	-0.862
L8	+0.0509	+0.00865	+8.19e-05	+2.91e-05	-0.0102	L26	+0.0168	+0.767	+0.00541	-0.00357	+0.0845
L9	+0.0566	+0.0166	+2.36e-05	+1.49e-05	-0.0587	L27	+0.238	+1.06	+0.00196	-0.0261	-0.139
L10	+0.0615	+0.0206	-4.16e-05	-5.96e-05	-0.106	L28	+0.172	+0.584	+0.00342	-0.013	+0.187
L11	+0.0342	+0.027	-0.000421	-0.000213	-0.0628	L29	+0.0245	+1.95	+0.000592	-0.0211	+0.0458
L12	+0.0476	+0.0292	-0.000179	-0.000235	-0.0722	L30	+0.062	+1.92	+0.00415	-0.0465	+0.0907
L13	+0.0577	+0.0673	-2.73e-06	-0.00159	-0.167	L31	-0.00269	+0.281	+0.00145	-0.0174	+0.0535
L14	+0.211	+0.167	-0.000527	-0.000391	-0.324	L32	+0.12	+1.97	-0.00194	-0.0207	+0.206
L15	+0.0807	+0.0424	-0.000514	-0.000319	-0.46	L33	+0.0727	+0.884	-0.00154	-0.0197	+0.265
L16	+0.101	+0.0511	-0.00101	-0.000685	-0.0977	L34	+0.198	+1.58	-0.00687	-0.0344	+0.236
L17	+0.306	+0.346	-0.00142	-0.00124	-0.251	L35	+0.136	+0.616	-0.00454	-0.0105	+0.374
L18	+0.451	+0.459	-0.000495	-0.000481	-0.233	L36	+0.0396	+0.014	+0.000329	-0.00028	-0.0882



Figure 27: Memory-source attention differences for Qwen3-8B: Middle importance minus baseline, per source token and averaged across heads, in units of 10^{-3}. All layers are shown.

	Target key	Target value	D key	D value	Final query key
L4	+0.0162	+0.0042	-0.000114	-0.000118	-0.00443
L8	+0.0593	+0.0055	+0.00917	+0.000342	+0.0195
L12	+0.121	+0.0755	+0.008	+0.00223	+0.0941
L16	+0.2	+0.0508	+0.00559	+0.00215	+0.129
L20	+0.714	+2.79	-0.00393	-0.00613	-1.28
L24	+0.125	+10.1	-0.0419	-0.613	+0.0128
L28	+0.107	+5.32	-0.00688	-0.308	+0.0198
L32	+0.205	+1.29	-0.00289	-0.0729	-0.241



Figure 28: Memory-source attention differences for Qwen3.5-9B: Middle importance minus baseline, per source token and averaged across heads, in units of 10^{-3}. Only the full/global attention layers listed in Table 22 are shown.

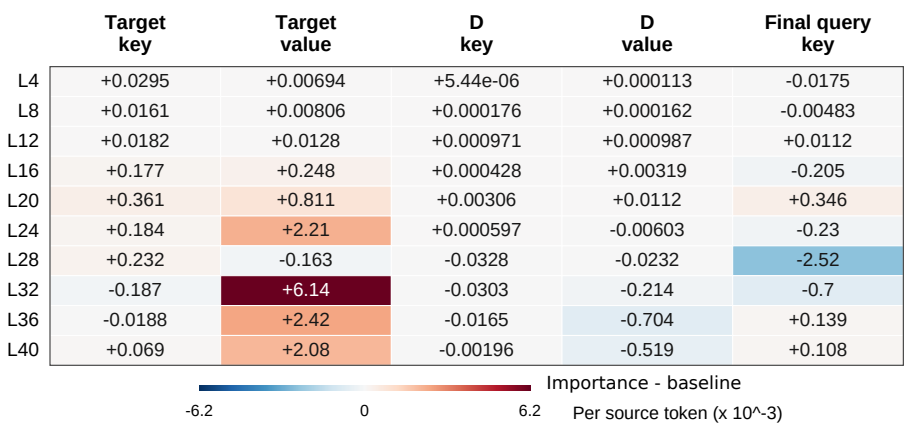


Figure 29: Memory-source attention differences for Qwen3.6-35B-A3B: Middle importance minus baseline, per source token and averaged across heads, in units of 10^{-3}. Only the full/global attention layers listed in Table 22 are shown.

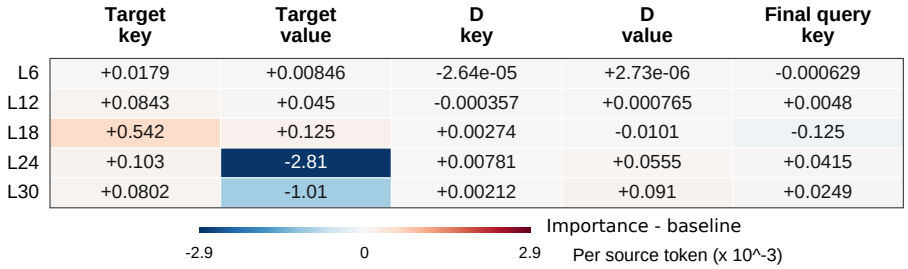


Figure 30: Memory-source attention differences for Gemma4-26B-A4B: Middle importance minus baseline, per source token and averaged across heads, in units of 10^{-3}. Only the full/global attention layers listed in Table 22 are shown.

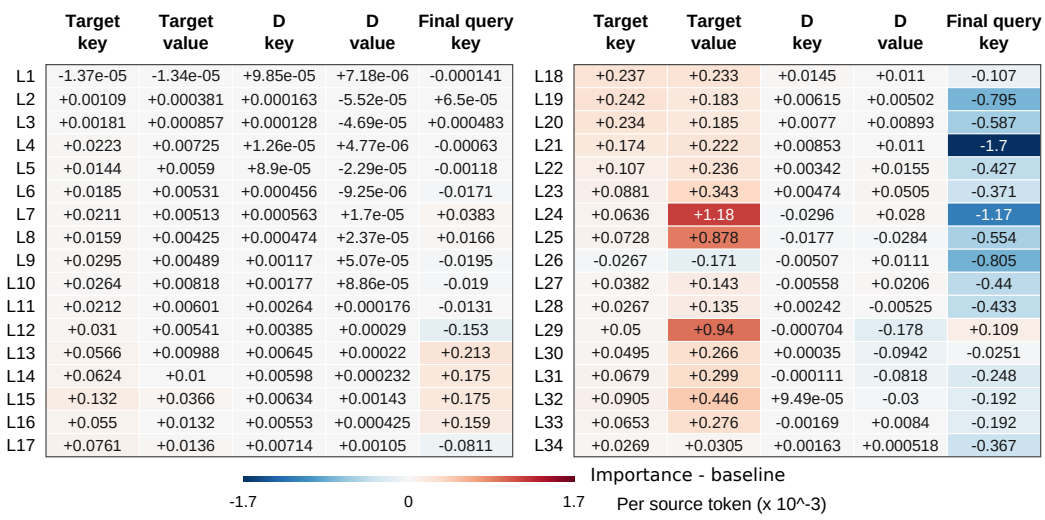


Figure 31: Memory-source attention differences for Ministral-3-8B: Middle importance minus baseline, per source token and averaged across heads, in units of 10^{-3}. All layers are shown.

F.3 MLP Geometry: A Cross-Model Descriptive Pattern

All five models reach their lowest sample-mean cosine in the latter half of the layer stack (Table 23). The minima occur at different depths: L26/36 for Qwen3-8B, L29/32 for Qwen3.5, L37/40 for Qwen3.6, L28/30 for Gemma4, and L32/34 for Ministral. Qwen3-8B and Qwen3.6 show deeper directional divergence, whereas the other three retain higher overall similarity. Several curves recover by the final layer. The depth and size of the directional change are model dependent, while its concentration in later layers is shared across models.

Table 23: Layer-wise summary of MLP geometry and shared steering. Cosine measures paired output-direction similarity; ΔP measures the change in target first-token probability after intervention.

Model	First cos.	Minimum cos.	Final cos.	Largest mean ΔP
Qwen3-8B	0.9841	0.7616 (L26)	0.9559	+0.0200 (L28)
Qwen3.5-9B	0.9999	0.9300 (L29)	0.9747	+0.0156 (L23)
Qwen3.6-35B-A3B	0.9999	0.7202 (L37)	0.9638	+0.0249 (L35)
Gemma4-26B-A4B	0.9999	0.9246 (L28)	0.9924	+0.0056 (L19)
Ministral-3-8B	1.0000	0.9297 (L32)	0.9669	+0.0209 (L28)

Sample heterogeneity. Qwen3-8B's first-layer mean cosine is 0.9841, compared with a median of approximately 0.99998 and a minimum of 0.5677. Thus, most first-layer outputs remain closely aligned while a smaller subset shows greater directional change. The complete layer tables show how this distribution evolves across depth through its mean, median, and observed range.

Relation to layer-wise retrieval dynamics. The late concentration of directional changes complements the stage identified by the value readouts. Gemma4 also shows late divergence, despite its opposite prompt-level effect. This motivates examining the transferred direction itself: a geometric change identifies where conditions differ, while steering tests how a shared component of that difference affects target support.

F.4 Shared MLP Steering and Complete Layer Profiles

The largest mean probability increases are 0.02003 at L28 for Qwen3-8B, 0.01557 at L23 for Qwen3.5, 0.02486 at L35 for Qwen3.6, and 0.02085 at L28 for Ministral. At these layers, target

probability increases in 69.5%, 84%, 94%, and 85% of held-out examples, respectively. The mean and median reveal variation in response magnitude: for Qwen3-8B at L28, the mean is near 0.02 while the median is 1.64×10^{-6}.

Gemma4's mean effects range from -0.00412 at L23 to $+0.00559$ at L19. Its effects are smaller than the positive peaks in the other four models and alternate in sign across layers. The maximum mean steering gain and minimum-cosine layer differ in every model: Qwen3-8B peaks after its cosine minimum, while the other models peak before it. The largest directional change and the strongest transferable effect therefore occur at distinct depths.

The following profiles pair each model's cosine curve with its layer-wise steering results. The tables report mean, median, minimum, and maximum cosine alongside mean and median probability changes and the percentage of held-out examples with positive changes.

F.4.1 QWEN3-8B: GEOMETRY AND HELD-OUT INTERVENTION

The paired-output cosine reaches its minimum at L26, followed by the largest mean steering gains at L28 and L29. Both analyses use 450-record contexts with 225 distractors and a target at index 225. Figure 32 shows both layer-wise profiles; Table 24 gives the complete layer statistics.

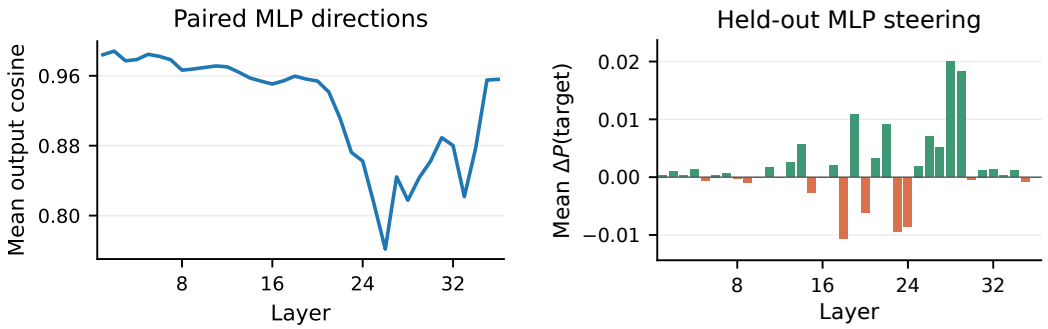


Figure 32: Qwen3-8B. Left: sample-mean paired MLP-output cosine for the 450-record, 225-distractor setting with the target at index 225. Right: mean target first-token probability change from a held-out single-layer intervention. Each summary contains 200 examples. Green/orange steering bars indicate increases/decreases in target probability.

Table 24: All 36 layers for Qwen3-8B. Mean/median/minimum/maximum cosine describe paired output directions. The final three columns describe held-out steering; positive fractions use the unrounded sign of ΔP.

Layer	Mean cos.	Median	Min.	Max.	Mean ΔP	Median ΔP	$\Delta P > 0$ (%)
1	0.9841	1.0000	0.5677	1.0000	+0.0004	0	42.0
2	0.9882	1.0000	0.6827	1.0000	+0.0010	0	46.5
3	0.9771	0.9998	0.3935	0.9999	+0.0003	0	42.5
4	0.9785	0.9997	0.4404	0.9999	+0.0014	0	47.5
5	0.9846	0.9998	0.5812	0.9999	-0.0007	0	42.0
6	0.9822	0.9997	0.5269	0.9998	+0.0004	0	47.0
7	0.9784	0.9997	0.4451	0.9998	+0.0006	0	47.0
8	0.9664	0.9993	0.1417	0.9995	-0.0002	0	45.5
9	0.9678	0.9992	0.1971	0.9995	-0.0010	0	45.0
10	0.9695	0.9988	0.2696	0.9992	-0.0001	0	43.5
11	0.9712	0.9985	0.3067	0.9990	+0.0017	8.8×10^{-14}	52.0
12	0.9703	0.9974	0.3032	0.9983	-0.0001	0	45.0
13	0.9643	0.9898	0.3346	0.9941	+0.0026	2.6×10^{-8}	60.0
14	0.9576	0.9858	0.2618	0.9923	+0.0056	1.5×10^{-10}	53.5
15	0.9539	0.9881	0.1523	0.9929	-0.0026	-6.9×10^{-14}	40.0
16	0.9506	0.9863	0.1389	0.9925	-0.0001	0	41.5
17	0.9543	0.9865	0.2336	0.9911	+0.0021	3.6×10^{-9}	58.5
18	0.9596	0.9881	0.3049	0.9922	-0.0107	-9.4×10^{-7}	18.5
19	0.9562	0.9873	0.2224	0.9917	+0.0109	1.1×10^{-6}	73.0
20	0.9540	0.9855	0.2116	0.9905	-0.0062	-1.8×10^{-7}	29.0
21	0.9414	0.9759	0.1320	0.9871	+0.0033	1.4×10^{-7}	56.5
22	0.9112	0.9395	0.2212	0.9728	+0.0092	4.3×10^{-10}	53.0
23	0.8723	0.9035	0.1294	0.9592	-0.0095	-8.3×10^{-7}	38.5
24	0.8624	0.8926	0.1296	0.9542	-0.0086	-1.6×10^{-8}	45.5
25	0.8135	0.8501	0.1432	0.9457	+0.0019	9.0×10^{-8}	54.0
26	0.7616	0.8012	0.0576	0.9256	+0.0070	6.8×10^{-8}	56.5
27	0.8443	0.8751	0.1520	0.9395	+0.0052	2.3×10^{-7}	61.0
28	0.8175	0.8506	0.1189	0.9333	+0.0200	1.6×10^{-6}	69.5
29	0.8433	0.8783	0.2042	0.9521	+0.0183	8.3×10^{-6}	82.0
30	0.8622	0.8882	0.3465	0.9745	-0.0004	-7.9×10^{-12}	41.5
31	0.8891	0.9104	0.4444	0.9848	+0.0012	1.3×10^{-7}	58.5
32	0.8802	0.9087	0.2625	0.9851	+0.0013	0	49.0
33	0.8217	0.8618	0.2138	0.9724	+0.0003	0	49.5
34	0.8771	0.9142	0.3005	0.9831	+0.0011	0	47.5
35	0.9552	0.9651	0.7813	0.9931	-0.0008	-1.2×10^{-7}	48.0
36	0.9559	0.9736	0.4027	0.9976	-0.0001	6.7×10^{-12}	50.5

F.4.2 QWEN3.5-9B: GEOMETRY AND HELD-OUT INTERVENTION

Directional similarity remains high in early layers and declines late. The strongest positive shared intervention precedes the minimum-cosine layer; the final-layer intervention is negative. Figure 33 shows both layer-wise profiles; Table 25 gives the complete layer statistics.

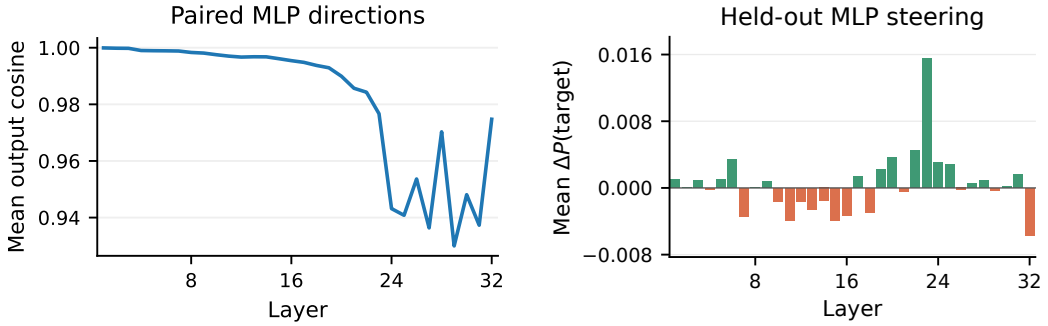


Figure 33: Qwen3.5-9B. Left: sample-mean paired MLP-output cosine. Right: mean target first-token probability change from a held-out single-layer intervention. Each summary contains 100 examples. Green/orange steering bars indicate increases/decreases in target probability.

Table 25: All 32 layers for Qwen3.5-9B. Mean/median/minimum/maximum cosine describe paired output directions. The final three columns describe held-out steering; positive fractions use the unrounded sign of ΔP.

Layer	Mean cos.	Median	Min.	Max.	Mean ΔP	Median ΔP	$\Delta P > 0$ (%)
1	0.9999	0.9999	0.9999	1.0000	+0.0010	+0.0002	59.0
2	0.9998	0.9998	0.9997	0.9999	1.7×10^{-5}	-2.2×10^{-6}	49.0
3	0.9998	0.9998	0.9994	0.9999	+0.0009	3.9×10^{-5}	54.0
4	0.9990	0.9990	0.9987	0.9992	-0.0002	1.4×10^{-5}	51.0
5	0.9989	0.9990	0.9987	0.9992	+0.0010	-0.0002	46.0
6	0.9989	0.9989	0.9986	0.9992	+0.0034	+0.0012	68.0
7	0.9988	0.9988	0.9985	0.9991	-0.0034	-0.0003	37.0
8	0.9983	0.9983	0.9979	0.9987	+0.0001	+0.0004	64.0
9	0.9981	0.9981	0.9975	0.9986	+0.0007	+0.0005	61.0
10	0.9975	0.9976	0.9967	0.9981	-0.0016	-0.0007	36.0
11	0.9970	0.9971	0.9962	0.9978	-0.0039	-0.0013	38.0
12	0.9967	0.9967	0.9956	0.9976	-0.0016	-0.0003	37.0
13	0.9968	0.9968	0.9959	0.9977	-0.0025	-0.0018	29.0
14	0.9968	0.9968	0.9957	0.9977	-0.0016	-0.0019	38.0
15	0.9961	0.9962	0.9943	0.9973	-0.0039	-0.0003	38.0
16	0.9954	0.9955	0.9935	0.9967	-0.0033	-0.0007	32.0
17	0.9948	0.9949	0.9922	0.9965	+0.0014	-0.0004	42.0
18	0.9938	0.9938	0.9906	0.9958	-0.0030	-1.2×10^{-5}	46.0
19	0.9929	0.9930	0.9889	0.9955	+0.0022	+0.0010	77.0
20	0.9899	0.9903	0.9833	0.9939	+0.0037	+0.0010	66.0
21	0.9857	0.9861	0.9728	0.9923	-0.0004	+0.0002	56.0
22	0.9843	0.9849	0.9698	0.9915	+0.0044	+0.0024	85.0
23	0.9767	0.9776	0.9617	0.9891	+0.0156	+0.0155	84.0
24	0.9432	0.9566	0.6844	0.9855	+0.0031	1.6×10^{-5}	52.0
25	0.9408	0.9559	0.7273	0.9810	+0.0028	+0.0005	67.0
26	0.9536	0.9602	0.8349	0.9856	-0.0002	+0.0006	62.0
27	0.9364	0.9519	0.8181	0.9864	+0.0006	+0.0003	64.0
28	0.9703	0.9759	0.8963	0.9949	+0.0009	+0.0012	73.0
29	0.9300	0.9489	0.6892	0.9893	-0.0003	-0.0006	33.0
30	0.9481	0.9571	0.8478	0.9898	+0.0001	1.9×10^{-6}	50.0
31	0.9373	0.9458	0.7403	0.9888	+0.0016	+0.0003	54.0
32	0.9747	0.9825	0.8695	0.9969	-0.0057	-0.0033	35.0

F.4.3 QWEN3.6-35B-A3B: GEOMETRY AND HELD-OUT INTERVENTION

The branch includes routed experts and the gated shared expert. A pronounced late cosine minimum coexists with both positive and negative neighboring intervention layers. Figure 34 shows both layer-wise profiles; Table 26 gives the complete layer statistics.

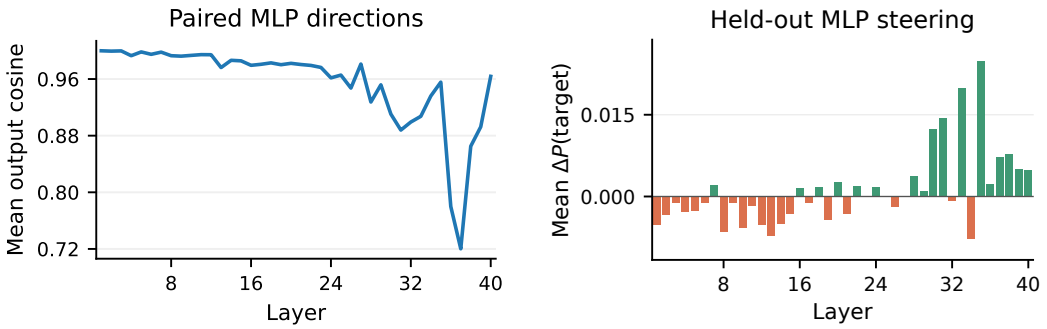


Figure 34: Qwen3.6-35B-A3B. Left: sample-mean paired MLP-output cosine. Right: mean target first-token probability change from a held-out single-layer intervention. Each summary contains 100 examples. Green/orange steering bars indicate increases/decreases in target probability.

Table 26: All 40 layers for Qwen3.6-35B-A3B. Mean/median/minimum/maximum cosine describe paired output directions. The final three columns describe held-out steering; positive fractions use the unrounded sign of ΔP.

Layer	Mean cos.	Median	Min.	Max.	Mean ΔP	Median ΔP	$\Delta P > 0$ (%)
1	0.9999	0.9999	0.9998	0.9999	-0.0052	-0.0014	44.0
2	0.9994	0.9996	0.9824	0.9997	-0.0034	-0.0010	42.0
3	0.9997	0.9998	0.9972	0.9998	-0.0012	-0.0005	46.0
4	0.9929	0.9992	0.9604	0.9995	-0.0028	-0.0009	47.0
5	0.9983	0.9992	0.9930	0.9995	-0.0027	-0.0008	43.0
6	0.9949	0.9992	0.9694	0.9996	-0.0012	-0.0015	43.0
7	0.9980	0.9991	0.9919	0.9994	+0.0022	-0.0001	49.0
8	0.9929	0.9989	0.9590	0.9994	-0.0064	-0.0010	47.0
9	0.9923	0.9977	0.9658	0.9989	-0.0012	-0.0004	48.0
10	0.9934	0.9969	0.9723	0.9991	-0.0058	-0.0027	42.0
11	0.9944	0.9962	0.9815	0.9986	-0.0017	1.9×10^{-5}	50.0
12	0.9943	0.9947	0.9802	0.9986	-0.0052	-0.0054	37.0
13	0.9762	0.9708	0.9457	0.9980	-0.0073	-0.0067	36.0
14	0.9864	0.9887	0.9680	0.9978	-0.0051	-0.0033	41.0
15	0.9857	0.9833	0.9736	0.9989	-0.0033	-0.0021	49.0
16	0.9794	0.9798	0.9566	0.9979	+0.0015	+0.0002	51.0
17	0.9808	0.9823	0.9615	0.9977	-0.0012	-0.0003	46.0
18	0.9828	0.9832	0.9695	0.9973	+0.0017	-0.0005	43.0
19	0.9802	0.9795	0.9621	0.9969	-0.0044	-0.0032	34.0
20	0.9822	0.9845	0.9635	0.9968	+0.0027	1.2×10^{-5}	51.0
21	0.9805	0.9824	0.9363	0.9937	-0.0033	-0.0010	43.0
22	0.9792	0.9825	0.9536	0.9955	+0.0018	-0.0002	49.0
23	0.9764	0.9774	0.9478	0.9963	+0.0001	-0.0010	46.0
24	0.9616	0.9619	0.9144	0.9922	+0.0018	+0.0002	54.0
25	0.9656	0.9663	0.9244	0.9901	+0.0001	-0.0001	50.0
26	0.9472	0.9498	0.8919	0.9843	-0.0018	-0.0033	44.0
27	0.9812	0.9839	0.9393	0.9946	-0.0002	-0.0002	45.0
28	0.9275	0.9427	0.7259	0.9748	+0.0038	+0.0026	57.0
29	0.9518	0.9596	0.8199	0.9881	+0.0011	+0.0003	51.0
30	0.9105	0.9245	0.6951	0.9811	+0.0124	+0.0055	74.0
31	0.8879	0.8990	0.6695	0.9778	+0.0143	+0.0138	71.0
32	0.8993	0.9075	0.7526	0.9609	-0.0008	-0.0003	49.0
33	0.9073	0.9151	0.7348	0.9705	+0.0199	+0.0166	80.0
34	0.9360	0.9380	0.8594	0.9740	-0.0077	-0.0060	29.0
35	0.9556	0.9625	0.8766	0.9915	+0.0249	+0.0243	94.0
36	0.7800	0.8077	0.3905	0.9647	+0.0022	+0.0017	60.0
37	0.7202	0.7656	0.1897	0.9667	+0.0073	+0.0048	79.0
38	0.8651	0.8861	0.6310	0.9797	+0.0079	+0.0040	72.0
39	0.8925	0.9206	0.6881	0.9855	+0.0051	+0.0040	70.0
40	0.9638	0.9732	0.7411	0.9909	+0.0049	+0.0001	51.0

F.4.4 GEMMA4-26B-A4B: GEOMETRY AND HELD-OUT INTERVENTION

Output-direction similarity declines late, while the shared steering effects remain small and alternate in sign. The largest mean probability increase is 0.00559 at L19, compared with a minimum cosine at L28. Figure 35 shows both layer-wise profiles; Table 27 gives the complete layer statistics.

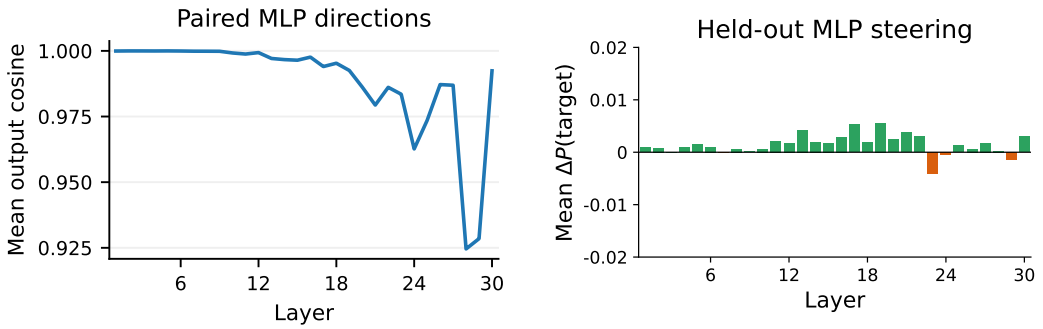


Figure 35: Gemma4-26B-A4B. Left: sample-mean paired MLP-output cosine. Right: mean target first-token probability change from a held-out single-layer intervention. Each summary contains 100 examples. Green/orange steering bars indicate increases/decreases in target probability.

Table 27: All 30 layers for Gemma4-26B-A4B. Mean/median/minimum/maximum cosine describe paired output directions. The final three columns describe held-out steering; positive fractions use the unrounded sign of ΔP.

Layer	Mean cos.	Median	Min.	Max.	Mean ΔP	Median ΔP	$\Delta P > 0$ (%)
1	0.9999	1.0000	0.9993	1.0000	+0.0009	5.7×10^{-7}	55.0
2	1.0000	1.0000	0.9998	1.0000	+0.0008	2.9×10^{-10}	50.0
3	1.0000	1.0000	0.9999	1.0000	1.4×10^{-5}	8.9×10^{-7}	57.0
4	1.0000	1.0000	0.9988	1.0000	+0.0009	6.8×10^{-8}	54.0
5	1.0000	1.0000	0.9999	1.0000	+0.0014	5.4×10^{-9}	50.0
6	0.9999	1.0000	0.9999	1.0000	+0.0010	5.5×10^{-7}	56.0
7	0.9999	0.9999	0.9997	0.9999	-0.0001	-4.1×10^{-9}	49.0
8	0.9999	0.9999	0.9993	1.0000	+0.0006	3.2×10^{-9}	51.0
9	0.9998	0.9999	0.9996	0.9999	+0.0002	-1.1×10^{-7}	46.0
10	0.9992	0.9994	0.9980	0.9997	+0.0006	3.5×10^{-7}	54.0
11	0.9988	0.9992	0.9949	0.9996	+0.0021	3.2×10^{-6}	57.0
12	0.9993	0.9994	0.9983	0.9997	+0.0017	3.1×10^{-8}	52.0
13	0.9971	0.9973	0.9938	0.9984	+0.0043	1.6×10^{-5}	66.0
14	0.9967	0.9969	0.9941	0.9978	+0.0020	8.0×10^{-6}	63.0
15	0.9964	0.9966	0.9939	0.9978	+0.0017	2.0×10^{-6}	62.0
16	0.9976	0.9977	0.9947	0.9986	+0.0029	1.5×10^{-5}	62.0
17	0.9940	0.9942	0.9900	0.9970	+0.0053	+0.0001	72.0
18	0.9953	0.9955	0.9899	0.9981	+0.0019	3.2×10^{-5}	64.0
19	0.9925	0.9927	0.9864	0.9964	+0.0056	+0.0002	77.0
20	0.9862	0.9868	0.9730	0.9935	+0.0025	2.5×10^{-5}	64.0
21	0.9794	0.9800	0.9635	0.9893	+0.0037	4.1×10^{-5}	68.0
22	0.9861	0.9866	0.9736	0.9936	+0.0030	8.3×10^{-6}	62.0
23	0.9835	0.9842	0.9651	0.9923	-0.0041	-1.7×10^{-8}	49.0
24	0.9627	0.9628	0.9204	0.9868	-0.0005	2.2×10^{-9}	51.0
25	0.9735	0.9746	0.9384	0.9895	+0.0012	2.9×10^{-6}	57.0
26	0.9872	0.9891	0.9646	0.9957	+0.0006	1.1×10^{-6}	58.0
27	0.9869	0.9892	0.9557	0.9977	+0.0018	8.4×10^{-6}	57.0
28	0.9246	0.9406	0.7236	0.9936	+0.0003	1.3×10^{-5}	60.0
29	0.9285	0.9431	0.7690	0.9929	-0.0014	2.2×10^{-5}	57.0
30	0.9924	0.9931	0.9742	0.9992	+0.0030	+0.0004	66.0

F.4.5 MINISTRAL-3-8B: GEOMETRY AND HELD-OUT INTERVENTION

The largest mean steering gain occurs at L28, followed by the minimum cosine at L32. Earlier layers show smaller changes, with both positive and negative effects. Figure 36 shows both layer-wise profiles; Table 28 gives the complete layer statistics.

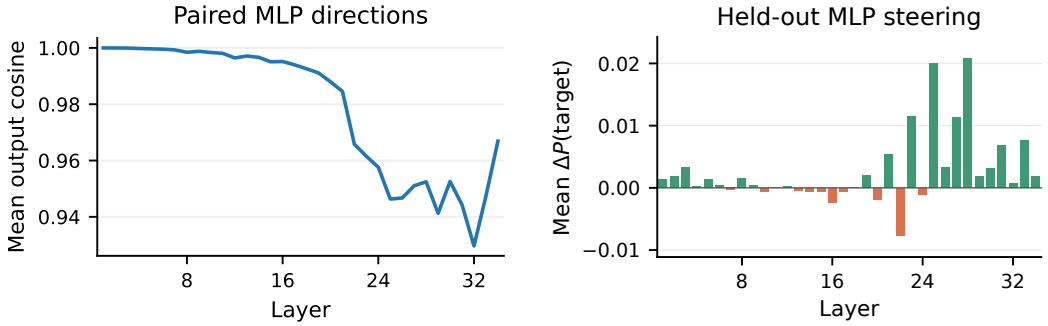


Figure 36: Ministral-3-8B. Left: sample-mean paired MLP-output cosine. Right: mean target first-token probability change from a held-out single-layer intervention. Each summary contains 100 examples. Green/orange steering bars indicate increases/decreases in target probability.

Table 28: All 34 layers for Ministral-3-8B. Mean/median/minimum/maximum cosine describe paired output directions. The final three columns describe held-out steering; positive fractions use the unrounded sign of ΔP.

Layer	Mean cos.	Median	Min.	Max.	Mean ΔP	Median ΔP	$\Delta P > 0$ (%)
1	1.0000	1.0000	1.0000	1.0000	+0.0013	+0.0001	54.0
2	1.0000	1.0000	0.9999	1.0000	+0.0018	+0.0001	54.0
3	0.9999	0.9999	0.9999	1.0000	+0.0032	+0.0001	59.0
4	0.9998	0.9998	0.9996	0.9999	+0.0002	-4.2×10^{-5}	49.0
5	0.9997	0.9997	0.9993	0.9998	+0.0014	+0.0003	59.0
6	0.9995	0.9996	0.9992	0.9997	+0.0004	2.3×10^{-5}	51.0
7	0.9993	0.9993	0.9987	0.9995	-0.0004	+0.0001	52.0
8	0.9984	0.9985	0.9979	0.9988	+0.0015	+0.0001	53.0
9	0.9988	0.9988	0.9984	0.9991	+0.0005	+0.0001	53.0
10	0.9984	0.9984	0.9980	0.9988	-0.0006	-1.9×10^{-5}	46.0
11	0.9981	0.9980	0.9976	0.9985	4.1×10^{-6}	-1.1×10^{-6}	50.0
12	0.9964	0.9965	0.9952	0.9972	+0.0003	-0.0003	42.0
13	0.9971	0.9971	0.9964	0.9978	-0.0004	-0.0002	39.0
14	0.9967	0.9966	0.9957	0.9974	-0.0006	-0.0004	38.0
15	0.9951	0.9950	0.9931	0.9962	-0.0006	-0.0002	41.0
16	0.9952	0.9952	0.9929	0.9966	-0.0023	-0.0007	37.0
17	0.9940	0.9940	0.9906	0.9960	-0.0006	+0.0006	61.0
18	0.9926	0.9928	0.9871	0.9954	1.8×10^{-5}	-0.0007	37.0
19	0.9911	0.9916	0.9859	0.9942	+0.0020	-1.4×10^{-6}	49.0
20	0.9880	0.9882	0.9813	0.9929	-0.0019	-0.0009	39.0
21	0.9846	0.9850	0.9776	0.9907	+0.0054	+0.0032	79.0
22	0.9658	0.9663	0.9379	0.9816	-0.0077	-0.0087	30.0
23	0.9615	0.9626	0.9199	0.9779	+0.0114	+0.0099	76.0
24	0.9576	0.9592	0.9244	0.9781	-0.0011	-0.0006	46.0
25	0.9463	0.9480	0.9008	0.9727	+0.0200	+0.0182	80.0
26	0.9467	0.9493	0.8902	0.9724	+0.0033	+0.0013	59.0
27	0.9511	0.9538	0.8944	0.9781	+0.0114	+0.0072	75.0
28	0.9525	0.9563	0.8720	0.9813	+0.0209	+0.0194	85.0
29	0.9413	0.9472	0.7842	0.9809	+0.0019	+0.0004	52.0
30	0.9526	0.9595	0.8551	0.9848	+0.0031	+0.0020	64.0
31	0.9443	0.9530	0.8042	0.9846	+0.0068	+0.0058	71.0
32	0.9297	0.9440	0.7698	0.9881	+0.0007	+0.0005	56.0
33	0.9474	0.9583	0.8409	0.9858	+0.0077	+0.0050	69.0
34	0.9669	0.9713	0.9169	0.9909	+0.0018	+0.0011	55.0