

What Survives When Activation Steering Is Used as Memory

Vedant Jumle

Delft University of Technology

vvjumle@student.tudelft.nl

Abstract

Activation steering has recently been used as a memory for skills: past experiences of an agent are stored as steering vectors and retrieved later. Facts and preferences about the user are assumed to need text instead. We test that assumption with a training-free activation memory beside a frozen LLM. The change an experience causes in late-layer states is stored as a shift, filed under a whitened key for the conversation’s gist, and added back in a later session only when the gist matches. We argue that such a shift is a fixed average of an effect that attention, in context, recomputes for every question, so only question-independent content can survive. In a pre-registered study on Qwen3.5-2B and 9B, with judges from four model families, the memory is selective (about 98% of unrelated replies are unchanged), but facts are stored as the word to say, in the last layers only, and are almost never used (at most 15% clean use, against 96–98% in context). Learnings transfer only in a narrow strength window. Most strikingly, the reference subtracted when writing a preference decides whether its concept or its direction survives: unless the opposite is subtracted, a stored dislike is recalled as a like.

1 Introduction

An assistant that is told “*I can’t stand jazz*” on Monday should not suggest a jazz playlist on Friday. Language models keep such information in one of two places. They keep it in their weights, which change only through training or model editing (Meng et al., 2022, 2023), or in their context, which is exact but is lost when the session ends unless it is retrieved and re-inserted (Lewis et al., 2020; Packer et al., 2023). A third place has recently been proposed: the model’s *activations*. Activation steering adds a vector to the residual stream at inference time and reliably shifts dispositions

such as honesty, refusal or persona (Turner et al., 2023; Zou et al., 2023; Rimskey et al., 2024; Lee et al., 2025; Cao et al., 2024). Zhao et al. (2026) go one step further and store an agent’s past experience as retrieved steering vectors, which works well as *procedural* memory, i.e. for skills. They explicitly leave *declarative* content (facts and preferences about the user) to text retrieval, on the assumption that it does not belong in activations.

That assumption has not been tested, and it matters for personalisation. Steering-based personalisation either keeps the user’s context in the prompt (He et al., 2025; Liu et al., 2024) or trains a vector for a behaviour chosen in advance (Cao et al., 2024). Neither shows what happens when what a user said in one conversation is stored as an activation change and replayed in a later conversation that no longer contains it.

We build such a memory and measure what survives. It is training-free and sits beside a frozen model (Figure 1a). When an experience occurs, the change it causes in late-layer states is stored as a *shift*, filed under a whitened *key* for the gist of the conversation. In a later session the shift is added back only where the current gist matches a stored key. Mathematically, the store is the linear associative memory that ROME and MEMIT write into MLP weights (Meng et al., 2022, 2023), kept in the activations instead, and the gate is a training-free variant of conditional steering (Lee et al., 2025).

Our account of what this memory can hold is simple. In context, attention recomputes an experience’s effect for each new question (Liu et al., 2024; Geva et al., 2023). A stored shift is one *fixed average* of that effect, so only the part of the experience that is the same for every later question can survive. This yields three predictions, which we test in a pre-registered study on Qwen3.5-2B and Qwen3.5-9B, with labels from a panel of LLM judges from different model families:

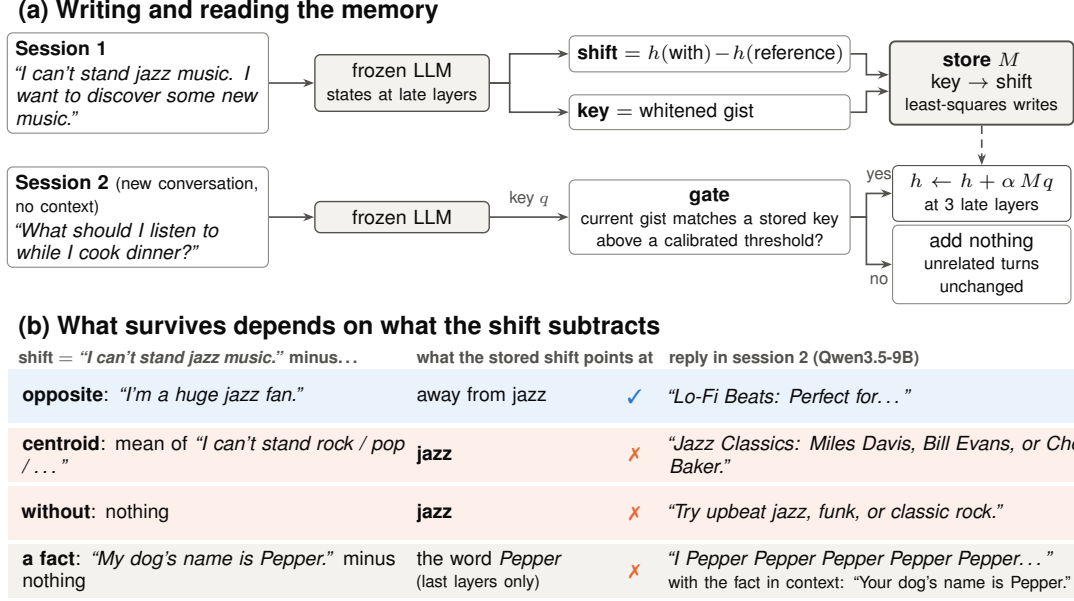


Figure 1: **(a)** A training-free activation memory beside a frozen LLM. In session 1 the change an experience causes in late-layer states is stored as a *shift*, filed under a whitened *key* for the conversation’s gist. In a later session the shift is added back only where the current gist matches a stored key. **(b)** What the stored shift carries depends on the reference subtracted when writing it. Subtracting the opposite keeps the *direction*; subtracting alternatives or nothing keeps the *concept*, so a dislike is recalled as a like. A fact is stored as the word to say. Replies are single real samples from Qwen3.5-9B (preferences $\alpha=1$, fact $\alpha=2$); quantitative results are in §5.

- **Facts fail.** A fact matters for one question, so its averaged shift reduces to “say the word”. The stored shift ranks the fact’s word in the top 10 of the vocabulary only in the last sixth of the network. Replies that use the fact cleanly stay at or below 15% at every strength, against 96–98% with the fact in context. Recall appears instead as loops and garbled text, and the memory never serves as a premise for yes/no questions.
- **Leanings survive only in a narrow window.** Broad preferences shift open-ended replies at moderate strength, but coherence collapses as strength grows. This claim failed its pre-registered test and is sensitive to the choice of judge.
- **What the shift subtracts decides what is stored.** A shift is the state with the experience minus a reference state. Subtracting the opposite experience keeps the preference’s *direction*. Subtracting a set of alternatives, or nothing, keeps its *concept*. Because negation is weakly encoded (Jang et al., 2023), a stored dislike is then recalled as a like: with the alternatives subtracted, the memory pushed every dislike item toward the thing the user

dislikes (original judge), and the average effect pointed the wrong way on both models under every judge that passed a sanity check (Figure 1b).

The memory is nevertheless selective: about 98% of unrelated replies are token-for-token identical to those with no memory, against 6–10% when the gate is forced open.

Our contributions are: (1) a gated, training-free activation memory for user content, and its link to associative-memory model editing; (2) the fixed-average account and the three predictions it makes; (3) a pre-registered test of these predictions on two model scales, with controls and a cross-family judge panel; and (4) the reference effect, a concrete way in which a naive activation memory inverts what the user said. Together, the results support the procedural/declarative split assumed by Zhao et al. (2026) and explain it: activation memory suits dispositions, while facts need weights or retrieval.

2 Method

We read and modify the residual stream $h_t^\ell \in \mathbb{R}^d$ at the output of a few late decoder blocks ℓ of a frozen model. Everything below is done separately at each of these layers.

Experiences and moments. An experience e (e.g. “*I can’t stand jazz music.*”) is presented with a few ordinary follow-up turns f (e.g. “*I want to discover some new music.*”). Each follow-up is one *moment*. We run each moment twice, once as e followed by f (*with*) and once with a reference text in place of e , and we compare the states over the tokens of f . Tokens of the chat template that follow the user’s text are excluded. In preliminary experiments, the choice of tokens to write did not matter (tokens selected by attention, by entropy or at random worked equally well), so a moment is stored as one pooled vector.

Shift and reference. The shift of moment j is a weighted mean over the follow-up’s tokens,

$$\Delta_j = \sum_{t \in f} w_t (h_t^{\text{with}} - h_t^{\text{ref}}), \quad w_t \propto H_t, \quad (1)$$

where H_t is the model’s next-token entropy at position t in the run without the experience, and the weights sum to one. The *reference* decides what Δ_j measures. We compare three:

- **without:** the follow-up alone, so Δ is everything the experience added;
- **opposite:** the opposite experience (“*I’m a huge jazz fan.*”), which cancels the shared topic and keeps the attitude;
- **centroid:** the mean state over eight same-frame alternatives (“*I can’t stand rock music.*”, “*... classical music.*”, ...), which cancels the frame and keeps what is specific to this experience. This is the construction used for concept vectors by Lindsey (2026).

Facts use *without*. Preferences use *opposite* unless stated otherwise.

Key. Let μ be the mean state on 100 generic prompts and their continuations, and let $W = U_r(\Lambda_r + \epsilon \bar{\lambda} I)^{-1/2}$ whiten the top $r=256$ principal directions of those states ($\epsilon=0.01$, $\bar{\lambda}$ the mean eigenvalue). The key of moment j is the normalised running mean of the whitened states over the follow-up’s user tokens in the run without the experience,

$$k_j = \text{normalize} \left(\frac{1}{|f|} \sum_{t \in f} W^\top (h_t - \mu) \right) \in \mathbb{R}^r. \quad (2)$$

The key thus describes the *situation* the experience was filed under, not the experience itself. Whitening suppresses the components that all states share; without it, keys of unrelated situations were nearly as similar as keys of related ones.

Store. A memory is a matrix $M \in \mathbb{R}^{d \times r}$ that maps keys to shifts. Its moments are written with the delta rule $M \leftarrow M + (\Delta_j - M k_j) k_j^\top$ of fast-weight memories (Schlag et al., 2021), so each moment’s own key recalls its shift exactly. When several memories share one store we use the least-squares solution (recursive least squares with ridge $\lambda=0.1$), which removes the dependence on write order. This is the closed-form key–value update of ROME and MEMIT (Meng et al., 2022, 2023), applied to activations rather than weights. In this paper each item has its own memory.

Read. In a later session, the read key q_t is the same pooled, whitened, normalised quantity over the user text seen so far. The gate opens when the best match with the stored keys exceeds a threshold θ ,

$$g_t = [\max_j \cos(q_t, k_j) > \theta], \quad (3)$$

$$h_t \leftarrow h_t + \alpha g_t M q_t, \quad (4)$$

where $[\cdot]$ is 1 if true and 0 otherwise, θ is the 95th percentile of the match on the generic prompts, set per item and layer. The gate is never open on the system prompt or the template head. Keys and gates are computed in a first, memory-free pass, so an injection cannot change its own gate; the second pass adds the recalled shift at each layer. The gate is a conditional-steering rule in the sense of Lee et al. (2025), except that the condition is written from a single experience rather than learned from labelled prompts, and its threshold is calibrated on ordinary text rather than tuned. α is the only free strength parameter, applied equally at each layer.

Layers. On Qwen3.5-2B (24 blocks) we read and write at block outputs 20, 21 and 23, chosen in a preliminary sweep over layers 4–23 on a subset of the items, in which three late layers together carried much more than any single layer (Appendix C). On Qwen3.5-9B (32 blocks) we use the same relative depths (27, 28, 31).

3 Why a stored shift keeps some content and not other

In context, an experience affects a later reply through attention. At each new position the model

attends back to the tokens of the experience and computes what it needs for the question at hand. [Liu et al. \(2024\)](#) make this explicit: the attention output is a mixture of the query’s own state and the context’s states, with weights that depend on the query. [Geva et al. \(2023\)](#) show the same for facts: the attribute is not present at the subject’s position in ready-made form; upper-layer attention extracts it, driven by the question.

A stored shift removes the recomputation. It is one vector per moment, averaged over the follow-up’s tokens, and it is added unchanged at every position where the gate is open. Whatever part of the experience’s effect varies with the question is averaged away; what is constant across questions remains. This gives three predictions.

P1: facts fail. A fact matters at one slot (“*your dog’s name is ____*”). Its question-specific retrieval does not survive averaging, and what remains is the already-extracted output, say “*Pepper*”, which only exists near the end of the network. Pushed at every position, such a shift either does nothing or makes the model repeat the word; it should not be used correctly in a reply, and it cannot act as a premise, which requires combining the fact with the question.

P2: leanings survive, in a dose window. A broad preference (“*I’m vegetarian*”) pushes in the same direction for any food question, so a constant push helps. But the push is added at every position whether or not it is needed, so beyond some strength it should overwhelm the reply.

P3: the reference decides what is stored. The shift is the state with the experience minus a reference state, so it keeps whatever differs between the two. Against the *opposite* experience the topic cancels and the attitude remains. Against *alternatives* in the same frame (“*I can’t stand rock / pop / ...*”) the frame cancels and the specific concept remains. Against *nothing*, both remain. Because negation is weakly represented ([Jang et al., 2023](#); [Lindsey, 2026](#)), the concept dominates the last two, so a stored *dislike* should be recalled as a *like*. For positively stated preferences, concept and direction agree, so all three references should push the same way.

4 Experimental setup

Models. Qwen3.5-2B (24 blocks) and Qwen3.5-9B (32 blocks), text path only, in fp32, with think-

ing disabled. The 2B ran on NVIDIA V100 (32 GB) and the 9B on A100 (80 GB) GPUs of the DelftBlue supercomputer ([Delft High Performance Computing Centre \(DHPC\), 2024](#)). Like the rest of the Qwen3.5 family ([Qwen Team, 2026](#)), both interleave linear-attention (Gated DeltaNet) and full-attention blocks.

Items. 41 items in four groups, each with one experience sentence, three follow-up turns that write the memory, and later prompts (Appendix A):

- **15 leanings** (e.g. vegetarian, tight budget, early riser), each with its opposite and eight same-frame alternatives;
- **8 dislikes** (e.g. “*I can’t stand jazz music.*”, “*I don’t drink alcohol.*”);
- **6 one-of-many** preferences (e.g. a home country, a favourite genre);
- **12 facts** (names, job, home town, favourite colour, ...).

Preference items are probed with 1 related and 5 ambiguous open-ended prompts (“*What should I listen to while I cook dinner?*”) and with balanced yes/no questions. Fact items are probed with a direct question, two *use* prompts that need the fact without asking for it (“*Write a short birthday message to my sister.*”), and balanced yes/no questions (“*Is my dog called Pepper?*” / “*... Max?*”). 50 unrelated prompts test leakage.

Conditions. No memory (nomem); the experience written into the prompt (ctx, the ceiling); the memory at $\alpha \in \{0.5, 1, 2, 3\}$ per layer; and four controls at $\alpha=2$: a random vector with the recall’s norm through the item’s own gate (rand), another item’s shift under this item’s gate (swap), a memory written from a neutral sentence (placebo), and the item’s memory with the gate removed (gate_on). For the mechanism (P3), the dislikes, one-of-many items and four leanings are also written with the *centroid* and *without* references at $\alpha \in \{0.5, 1, 2\}$.

Generation. Each prompt gets one greedy and four sampled replies ($T=1$, top- p 0.95, top- k 20, presence penalty 1.5, at most 120 tokens). Every condition uses the same random streams, so differences come from the memory. A second seed set repeats the main conditions.

Measures. Replies are labelled by an LLM judge (Zheng et al., 2023) (Qwen3.5-9B, thinking off) with fixed rubrics (Appendix B). For preferences the primary measure is the *judge lean*: +1 for a coherent reply whose recommendations fit the user’s trait, −1 for a coherent reply against it, and 0 otherwise. Incoherent replies count as 0, so loops and garbled text never count as success. For facts the primary measure is *clean use* on the use prompts: the reply uses the correct value as a fact about the user and is coherent. We also report word-list measures, repetition loops (repeated 4-gram rate ≥ 0.3), the judge’s incoherence rate, yes/no accuracy and margins, and the logit lens (nostalgebraist, 2020) of stored shifts. Every number has a 95% bootstrap confidence interval over items and prompts, paired across conditions.

Pre-registration and judge robustness. Claims, decision rules, an item-inclusion rule (an item is kept if its in-context effect is at least 0.3) and the statistics were committed before any run (Appendix E). Because the original judge is from the same family as the models it grades, and LLM judges can favour their own outputs (Zheng et al., 2023), we re-labelled all 34,200 replies with three judges from other families (gpt-oss-20B, Gemma-4-12B, Llama-3.1-8B) and a larger Qwen model, and recomputed every judge-based verdict under each judge and under majority votes (Appendix B). This panel was added after the pre-registered analysis.

5 Results

5.1 The memory is selective

With the gate, 98.4% (2B) and 98.2% (9B) of replies to the 50 unrelated prompts are token-for-token identical to those with no memory, at $\alpha=2$. With the gate removed, only 5.7% and 10.1% are; the KL divergence from the no-memory distribution rises four- to six-fold (Figure 6 in the appendix). The gate opens somewhere on 3.5% and 4.0% of unrelated prompts. Selectivity therefore does not explain any of the failures below: when the memory misbehaves, it does so where it is supposed to fire.

5.2 Facts are stored as a late word and are not used

What is stored. Read through the output layer, the stored fact shift ranks the fact’s word in the top 10 of the vocabulary only at relative depth ≥ 0.83

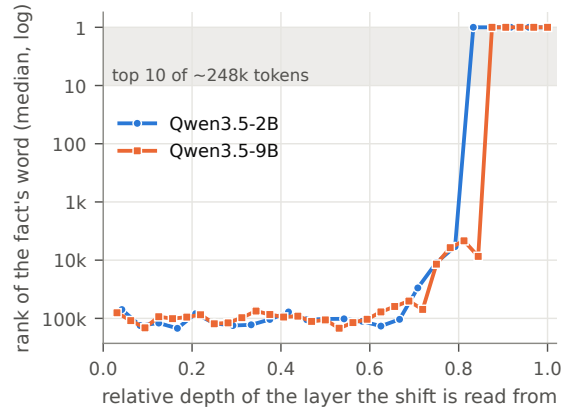


Figure 2: The stored fact shift read through the output layer at every block (logit lens). The median rank of the fact’s word (e.g. *Pepper*) among about 248k tokens stays near 100k until the last sixth of the network, then jumps to the top.

(2B) and ≥ 0.875 (9B); at half depth its median rank is about 70,000 on both models (Figure 2). The shift encodes the word to say, and only where saying it is decided. This was pre-registered as C6 and holds on both models. The logit lens is only approximate before the last layers (Belrose et al., 2023), but the behavioural results below point the same way.

What happens. Clean use of the fact stays far below the in-context ceiling at every strength (Figure 3): at most 0.15 (2B) and 0.12 (9B), at $\alpha=1$, against 0.96 and 0.98 in context. Increasing the strength does not convert mentions into use. At $\alpha=2$, 89% (2B) and 86% (9B) of replies mention the fact without using it cleanly, mostly as loops (82% on both). A typical reply to “*What’s my dog’s name?*” at $\alpha=2$ is “*I Pepper Pepper Pepper...*”; in context it is “*Your dog’s name is Pepper.*” At lower strength the model usually answers that it does not know. The three controls never produce the target. This was pre-registered as C3 and holds on both models and under every judge that passed our sanity check (Appendix B).

No premises. On balanced yes/no questions about the facts, the models without memory answer “No” to everything (balanced accuracy 0.50); in context they reach 0.94. The memory never improves the balanced accuracy at any strength, on either model. On the 2B, at $\alpha=2$, it even lowers it (by 0.15, CI $[-0.27, -0.04]$), which formally fails the pre-registered rule for C4, whose wording required no change. The yes/no margins move by a

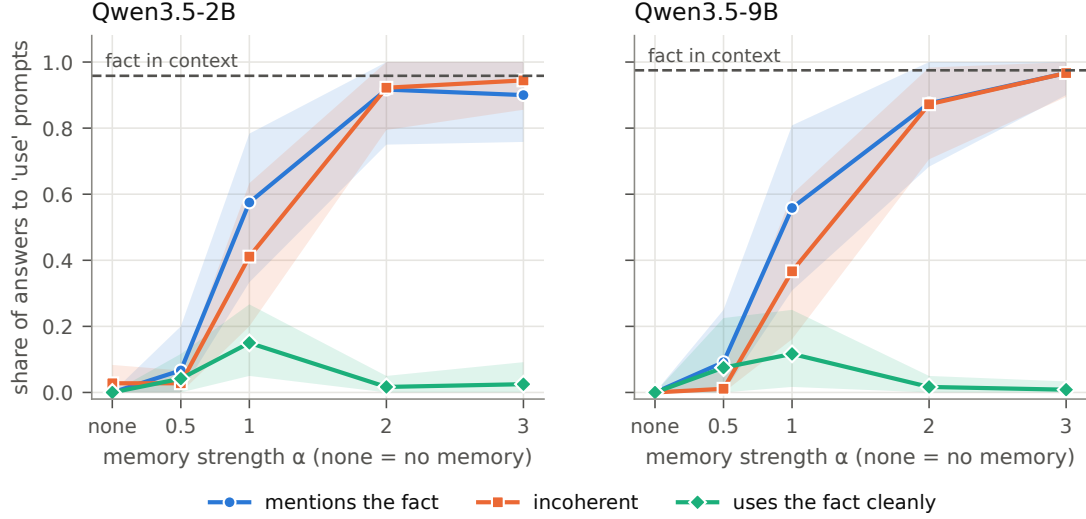


Figure 3: Fact memories on the use prompts. As the strength grows, more replies mention the fact (blue) and more are incoherent (orange), but replies that use the fact cleanly (green, judge) stay at or below 15%. With the fact in context, 96–98% use it cleanly (dashed). Shaded: 95% CIs over items and prompts.

general tilt rather than toward the correct answer: on the 9B at $\alpha=2$ the Yes-correct margin rises by 0.78 and the No-correct margin falls by 0.73. The same holds for the preference items. We read this as no evidence for premises on either model, with the 2B result being a decrease rather than a gain.

5.3 Leanings transfer only in a narrow window

At $\alpha=1$, leaning memories shift open-ended replies in the right direction: the judge lean rises by 0.20 (CI [0.04, 0.38], 2B) and 0.11 ([0.004, 0.24], 9B) over no memory, against 0.72 and 0.81 for the experience in context (Figure 4). An independent second set of samples reproduces these values within about 0.03. At $\alpha=2$ the word-list lean keeps rising, but the judge calls 61% (2B) and 41% (9B) of replies incoherent, against 14% and 3% without memory, and the gain’s interval includes zero. At $\alpha=3$ most replies are incoherent. The random-vector control’s gain is near zero, but its paired difference from the memory at $\alpha=2$ also includes zero. The pre-registered claim C2, which required a gain at both $\alpha=1$ and $\alpha=2$ and a clear separation from every control, therefore *fails* on both models.

This failure depends on the judge. The original judge is the strictest about coherence: it labels 62% of preference replies coherent, against 72–76% for the other competent judges. Under a majority vote of the three cross-family judges, the $\alpha=2$ gain is 0.36 ([0.15, 0.58], 2B) and 0.28 ([0.10, 0.48], 9B),

and the controls separate. The effect is also uneven across items: strong for *early riser*, *prefers quiet*, *vegan* and *vegetarian*, absent for *celiac*, *lactose intolerant* and *has young kids*. We conclude that leanings do transfer, but only within a narrow strength window, with a size that depends on how strictly incoherence is judged. Single-item preferences (one-of-many) barely move at any strength (gain at most 0.14).

5.4 What the shift subtracts decides concept or direction

For dislikes, the three references move replies in opposite directions (Figure 5). Compared at the same loop rate, the memory written against the **opposite** moves replies away from the disliked thing, by +0.23 ([0.05, 0.44], 2B) and +0.25 ([0.02, 0.50], 9B). The memory written against the **centroid** moves them toward it, by −0.38 ([−0.57, −0.19]) and −0.37 ([−0.59, −0.15]), and does so for every item on both models. The memory written against **nothing** also moves them toward it, by −0.25 and −0.31, for 5 of 7 and 7 of 8 items. Asked for music to cook to after “*I can’t stand jazz music.*”, the 9B with the centroid memory recommends “*Jazz Classics: Miles Davis, Bill Evans, or Chet Baker*” (Figure 1b). This was pre-registered as C5 and holds on both models. It also holds under gpt-oss, Gemma and both majority votes, with larger effects (−0.42 to −0.53 for the centroid). The same-family Qwen3.6 judge agrees on every sign, but its

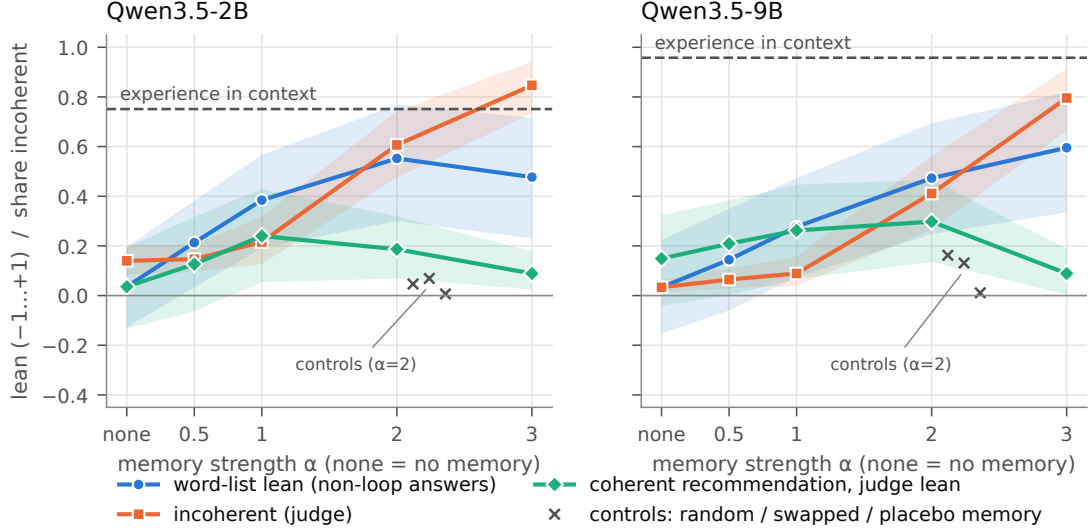


Figure 4: Leaning memories (15 items) under the original judge. The word-list lean of non-looping replies (blue) keeps rising with strength, but the share of incoherent replies (orange) rises faster, so the judge lean of coherent recommendations (green) peaks at $\alpha \approx 1-2$ and falls. Crosses: controls at $\alpha=2$. Dashed: the experience in context.

without interval touches zero.

The logit lens shows the cause directly (Figure 7 in the appendix). For dislikes, the stored shift’s preference for liked-side over disliked-side words at the last layer is +3.7 (2B) and +4.2 (9B) for the opposite reference, but −10.6 and −11.6 for the centroid and −3.2 and −3.6 for *without*: those shifts point at the disliked thing. For positively stated leanings, all three references point the same way (+4.8 to +9.1), and their behavioural effects agree in sign, as P3 predicts.

6 Related work

Steering and its limits. Adding vectors to the residual stream controls dispositions such as honesty, refusal and persona (Turner et al., 2023; Zou et al., 2023; Rimsky et al., 2024; Cao et al., 2024). Its effects vary across prompts and concepts (Tan et al., 2024; Braun, 2025), suit traits that shape a whole reply better than local ones (Subbiah et al., 2026), and trade expression for coherence (Subbiah et al., 2026; Suri et al., 2025). Injected concepts overwhelm the model at high strength, are read as the model’s own, and are not removed by negation (Lindsey, 2026). Our leaning and fact results are consistent with all of these; we add the memory setting and the role of the reference.

Conditional and personalised steering. CAST gates a steering vector on the similarity between the current state and a learned condition (Lee et al.,

2025); our gate is the same idea with the condition written from a single experience. In-context vectors compress demonstrations into a vector (Liu et al., 2024), and context steering amplifies the effect of a user’s context on the output probabilities (He et al., 2025); both need the context present. BiPO trains a personalised steering vector for a chosen behaviour (Cao et al., 2024). Context steering notes, informally, that a single context also makes its topic more likely and that a contrasting pair isolates the attribute (He et al., 2025); we show the consequence for stored memories systematically.

Memory in activations and weights. Neural Procedural Memory stores an agent’s skills as retrieved steering vectors and leaves declarative content to text (Zhao et al., 2026); our results support that division and explain it. Facts can be written usefully into weights by key-value updates at the subject’s position in middle-layer MLPs (Meng et al., 2022, 2023); our store uses the same update in activations, without optimising the stored value. Factual recall itself relies on question-driven attention in upper layers (Geva et al., 2023), which a stored shift cannot reproduce. Retrieval-augmented generation and agent memory systems keep facts in context (Lewis et al., 2020; Packer et al., 2023). Our store is also a fast-weight memory, the same delta-rule mechanism that linear-attention layers implement (Schlag et al., 2021).

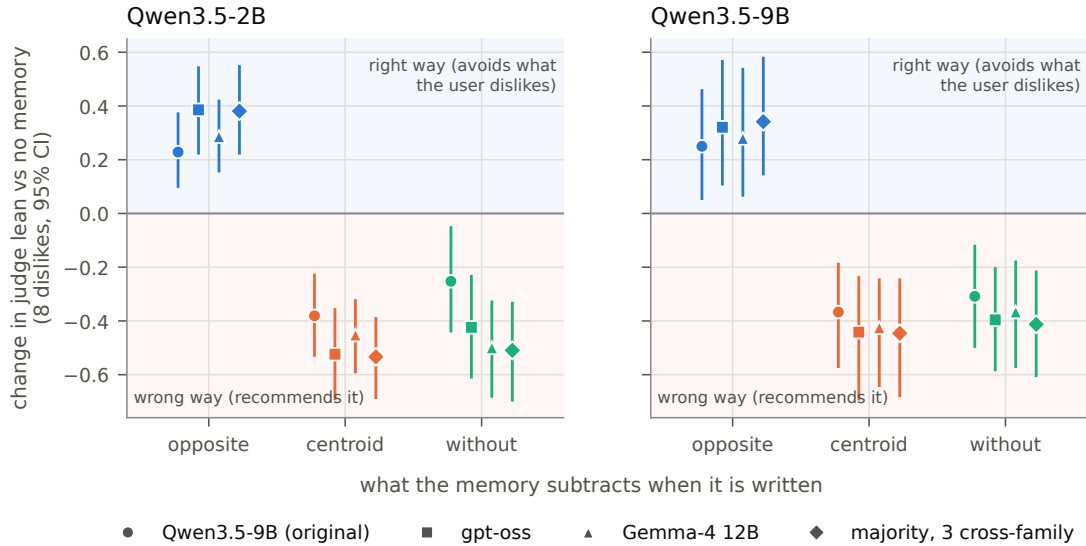


Figure 5: Dislike memories (2B: 7 items kept by the inclusion rule; 9B: 8), written with each reference and compared at the highest strength with at most 10% loops ($\alpha=1$ for all). Change in judge lean against no memory: positive means the replies avoid what the user dislikes. Only the opposite reference moves replies the right way; the centroid and *without* references push toward the disliked thing, under every competent judge. Bars: 95% CIs.

7 Conclusion

A gated activation memory beside a frozen model can carry broad leanings into a later session, selectively and only within a narrow strength window. It cannot carry facts or premises: a stored fact is the word to say, present only at the end of the network. And the way the memory is written determines what it means: unless the opposite experience is subtracted, a stored dislike is recalled as a like. We read all three results as consequences of one fact: a stored shift is a fixed average of an effect that attention recomputes, in context, for each question. Activation memory is therefore a natural home for dispositions, while facts need weights or retrieval. A practical next step is to store the concept and its sign separately, so that the direction of a preference no longer depends on a hand-written opposite.

Limitations

Scale and items. Two models from one family (2B and 9B) and 41 items, with 6–15 per group. The dislike result rests on 7–8 items. Larger models or other families may behave differently.

Judges. The primary labels come from an LLM judge from the same family as the graded models. The cross-family panel shows that C3 and C5 do not depend on the judge, but not that the labels are correct; no human labels were collected. Llama-3.1-8B was excluded from the panel after the fact

because it failed an objective check (Appendix B); its verdicts are reported. The panel itself was not pre-registered.

Design. Each memory is written from three hand-written follow-ups, and the opposite and centroid references are hand-written. A practical system would need these references without human input. We test isolated memories only; with many memories in one store, interference may change the picture. The theoretical account in Section 3 is an interpretation that fits the results, not a proof.

Ethical Considerations

The memory studied here stores information about users inside a model’s activations. Such stores raise the same privacy questions as any user memory (consent, inspection and deletion), with the added difficulty that an activation shift is not human-readable. Our results also show two failure modes that matter for deployment: a naive memory can recall a dislike as a like, and an injected memory can be presented by the model as its own trait. All experiments use synthetic user statements written by the author; no personal data was collected.

Acknowledgements

Experiment code, analysis scripts and the drafting of this paper were assisted by Claude (Anthropic). The author designed the study and the

pre-registration, and checked all reported results. The author acknowledges the use of computational resources of the DelftBlue supercomputer, provided by Delft High Performance Computing Centre (<https://www.tudelft.nl/dhpc>).

References

- Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.
- Joschka Braun. 2025. Understanding unreliability of steering vectors in language models. Master’s thesis, University of Tübingen. A version was presented at the ICLR 2025 Workshop on Foundation Models in the Wild.
- Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *arXiv preprint arXiv:2406.00045*.
- Delft High Performance Computing Centre (DHPC). 2024. DelftBlue Supercomputer (Phase 2). <https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2>.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Jerry Zhi-Yang He, Sashrika Pandey, Mariah L. Schrum, and Anca Dragan. 2025. Context steering: Controllable personalization at inference time. In *The Thirteenth International Conference on Learning Representations*.
- Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Proceedings of the 1st Transfer Learning for Natural Language Processing Workshop*, volume 203 of *Proceedings of Machine Learning Research*. PMLR.
- Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehl, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*.
- 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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.
- Jack Lindsey. 2026. Emergent introspective awareness in large language models. *arXiv preprint arXiv:2601.01828*. Also published in the Transformer Circuits Thread, <https://transformer-circuits.pub>.
- Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.
- Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.
- nostalgebraist. 2020. interpreting GPT: the logit lens. LessWrong blog post.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*.
- Qwen Team. 2026. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.
- Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. 2021. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning*.
- Melanie Subbiah, Zara Hall, and Kathleen McKeown. 2026. On the limits of steering vectors for preference-aligned generation. *arXiv preprint arXiv:2607.01802*.
- Manan Suri, Nishit Anand, and Amisha Bhaskar. 2025. Mitigating memorization in LLMs using activation steering. *arXiv preprint arXiv:2503.06040*.
- Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Chengfeng Zhao, Yuqiao Tan, Shizhu He, Yequan Wang, Jun Zhao, and Kang Liu. 2026. Neural procedural memory: Empowering LLM agents with implicit activation steering. *arXiv preprint arXiv:2606.29824*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang 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, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Items and prompts

Group	Items
Leanings (15)	vegetarian, vegan, lactose intolerant, celiac, diabetes, loves hiking, tight budget, early riser, prefers quiet, eco-conscious, has young kids, loves cooking, loves cold, bookworm, history buff
Dislikes (8)	jazz, alcohol, horror films, spicy food, hiking, seafood, the beach, sweet food
One-of-many (6)	loves jazz, lives in Norway, lives in Singapore, Japanese food, tennis, horror movies
Facts (12)	dog’s, cat’s, sister’s, brother’s, grandmother’s, best friend’s and boss’s names; job; favourite colour; home town; first language; university

Table 1: The 41 items. Each has an experience sentence (e.g. “*I can’t stand jazz music.*”), an opposite (“*I’m a huge jazz fan.*”), three follow-up turns and its probes; preferences also have eight same-frame alternatives for the centroid. The pre-registered inclusion rule removed one item (*hates horror*) on the 2B.

B Judges

Rubrics. The preference rubric asks whether the reply’s recommendations fit the user’s trait (*toward*), go against it (*away*) or neither (*neutral*), explicitly judging what is recommended rather than which words are mentioned; whether the reply is coherent; and whether the assistant claims the trait as its own. The fact rubric asks whether the reply uses the correct value as a fact about the user (a mention inside garbled or repetitive text does not count), whether it is coherent, whether the assistant claims the fact as its own, and whether it asserts

a wrong value. Full prompts are in <https://github.com/vedant-jumle/Seahorse>.

Panel. All 34,200 replies were re-labelled by gpt-oss-20B, Gemma-4-12B, Llama-3.1-8B and Qwen3.6-35B-A3B (same family; reported, not voting), with the same prompts, temperature 0 and JSON-constrained output; fewer than 0.1% of outputs failed to parse. Table 2 gives each claim’s verdict under each judge.

Sanity check. A reply cannot use a name it does not contain. Among replies a judge labelled as using the fact, the target string was absent in 1% for Qwen3.6, Gemma and gpt-oss, but in 48% for Llama-3.1-8B, which labels 92% of all fact replies as using the fact (others: 16–26%) and agrees with the other judges on this label at chance level ($\kappa \approx 0.05$). We therefore exclude it. Among the remaining judges, agreement on coherence is high (Fleiss’ $\kappa = 0.90$ on facts) and the original judge agrees with Gemma on fact use at $\kappa = 0.93$.

C Preliminary experiments

The design choices were fixed in earlier, smaller experiments that were not pre-registered. On Qwen2.5-1.5B, unwhitened per-token keys made the memory fire on unrelated prompts; whitened pooled keys with a calibrated threshold made unrelated replies identical to no memory. Selecting which tokens to write by attention, by entropy or at random made no difference. On Qwen3.5-2B, a sweep over layers 4–23 found fact-specific effects only in the last five layers, and three late layers together carried much more than any single one. An earlier version of the reference comparison (11 preference items) first showed the dislike reversal that core experiment C5 tests.

D Additional figures

E Pre-registration and reproduction

The claims (C1 selectivity, C2 leanings, C3 facts, C4 no premises, C5 reference, C6 a late word), the primary measures, the decision rules, the item-inclusion rule and the bootstrap were committed before any run (experiments/core_v1/PREREG.md). Results against the pre-registered rules: C1, C3, C5 and C6 hold on both models; C2 fails on both; C4 holds on the 9B and fails on the 2B (a decrease in accuracy, not a gain). Deviations: the

Claim	Model	Original	gpt-oss	Gemma	Llama [†]	Qwen3.6 [‡]	Maj. (orig.+3)	Maj. (3 cross)
C2 leanings	2B	no	no	no	no	yes	yes	yes
	9B	no	no	no	no	no	yes	yes
C3 facts	2B	yes	yes	yes	no	yes	yes	yes
	9B	yes	yes	yes	no	yes	yes	yes
C5 dislikes	2B	yes	yes	yes	no	no	yes	yes
	9B	yes	yes	yes	no	no	yes	yes

Table 2: Does each pre-registered claim hold under each judge? [†]Fails the sanity check below and is excluded from our conclusions. [‡]Same family as the original judge; its C5 signs all agree, but its *without* interval touches zero.

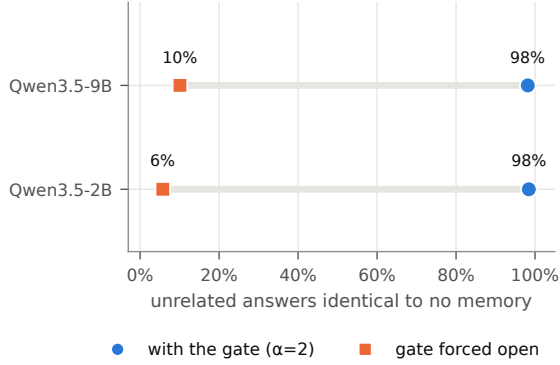


Figure 6: Unrelated replies identical to no memory, at $\alpha=2$, with the gate and with the gate forced open.

judge panel and the exclusion of Llama-3.1-8B were added after the pre-registered analysis; the pre-registered human audit of the judge was not carried out. All code, configurations, seeds and the commands that regenerate every table and figure from the saved outputs are available in <https://github.com/vedant-jumle/Seahorse>.

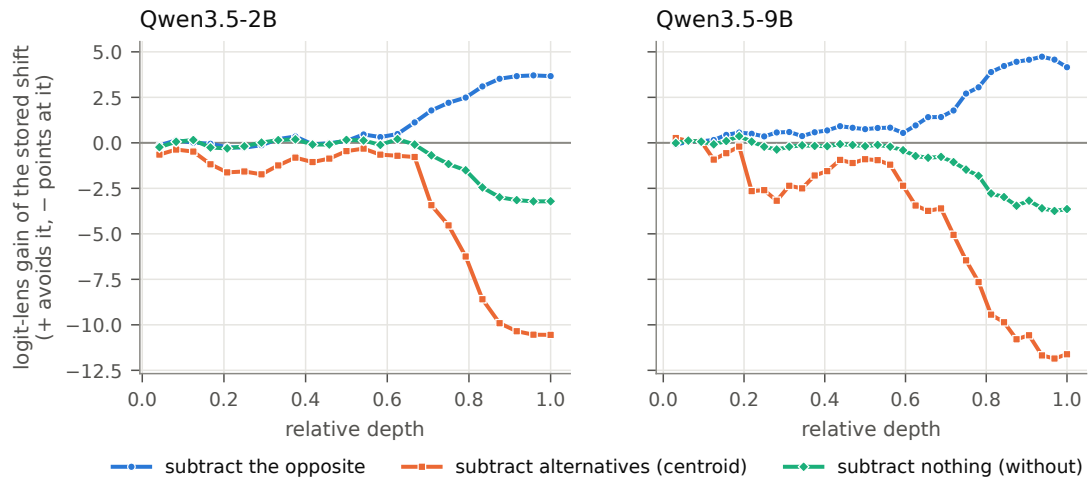


Figure 7: Logit lens of the stored dislike shifts at every block: mean logit of words on the liked side minus the disliked side (positive: avoids the disliked thing). The centroid and *without* shifts point at the disliked thing in the late layers.