On the Importance of Gating

Memorization vs. In-Context Learning in State Space Models

Paper by William L. Tong, Aryo Lotfi, Emmanuel Abbe,
Kostas Vaggelakos, Vishnu Banna, Etai Littwin, Josh Susskind,
Cengiz Pehlevan, and Eran Malach

Apple / Harvard University

Independent paper-reading presentation

arXiv:2609.16540v1 | 28 September 2026

Gating changes both which solution is learned and which history survives.

Two questions explain the apparent contradiction

Learning a retrieval solution

A correct solution exists, but strong decay weakens its gradient signal.

The model can first learn a stable mapping stored in its weights.

Using it in a longer context

Retaining every token also retains interference from irrelevant keys.

Decay can help when relevant information stays close to the query.

Weak gating helps discovery; selective forgetting can help extrapolation.

Paper, Secs. 1–2, PDF pp. 1–5. The second claim requires locality.

Perturbed labels make memorization distinguishable

- Fix P keys $x_i \in \mathbb{R}^d$ and base labels $y(x_i) \in \{-1, +1\}$.
- Resample a flip on each context occurrence:

$$\widetilde{y}_i = \eta_i y(x_i), \qquad \Pr(\eta_i = -1) = \epsilon.$$

- Query a key that appears in the context; predict its current label.

Illustrative example

A familiar key usually means $+1$. This prompt assigns it -1.
A weight lookup returns $+1$; successful context retrieval returns -1.

At $\epsilon = 0.1$, memorizing base labels gives 90% single-label accuracy.

Paper, Sec. 2.1, PDF pp. 2–3. Example is a teaching illustration.

A scalar decay controls how much history reaches the query

$$\widetilde{x}_t = (x_t, \widetilde{y}_t), \quad \widetilde{x}_q = (x_q, 0), \quad h = \widetilde{x}_q + \sum_{t=1}^{\ell} e^{-(\ell-t+1)e^{\lambda}} (\widetilde{x}_q^{\top} \widetilde{x}_t) \widetilde{x}_t$$

$$f = w^{\top} h, \qquad w = (w_x, v) \in \mathbb{R}^{d+1}.$$

- Train only the readout w and global log-decay rate λ.
- Initialize $w = 0$; use logistic loss $\log(1 + e^{-\widetilde{y} f})$.
- At distance Δ:

$$\alpha_{\Delta} = e^{-\Delta e^{\lambda}} = e^{-\Delta/\tau}, \qquad \tau = e^{-\lambda}.$$

Larger λ means stronger decay, smaller α_{Δ}, and shorter τ.

Paper, Sec. 2.2, PDF p. 3. A simplified Mamba-like layer with identity Q/K/V.

Two margins reveal which solution the model uses

Assume orthonormal keys and fix the matching position.

$$u = y(x_i)(w_x^\top x_i), \qquad \delta_m = (1+\alpha)u, \qquad \delta_r = \alpha v.$$

Signed classification margin

$$z = \widetilde{y} f = \eta \delta_m + \delta_r$$

Unflipped label: $\delta_m + \delta_r$.
Both mechanisms help.

Flipped label: $-\delta_m + \delta_r$.
Retrieval must overcome memorization.

$w_x = 0,\ v > 0$ already solves this idealized classification task.

Paper, Sec. 2.3, PDF p. 3; App. B.1, PDF p. 15.

Strong decay suppresses the initial retrieval gradient twice

At zero readout initialization,

$$\dot{\delta}_m(0) = \frac{(1+\alpha)^2(1-2\epsilon)}{2P}, \qquad \dot{\delta}_r(0) = \frac{\alpha^2}{2}.$$

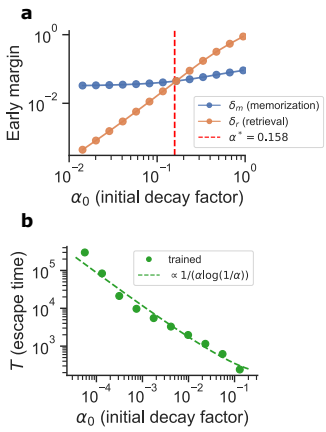
Proposition 1: initial growth-rate ratio

$$\frac{\dot{\delta}_r(0)}{\dot{\delta}_m(0)} = \frac{P\alpha^2}{(1-2\epsilon)(1+\alpha)^2}.$$

- One factor α weakens learning of v; another weakens its output.
- Shared retrieval benefits from all keys; each memory direction sees $1/P$.
- For small α, the crossover scale is $\alpha^* \approx \sqrt{(1-2\epsilon)/P}$.

Paper, Proposition 1 / Eq. (B.1), PDF p. 15. An initial-rate result, not a global phase transition.

A memorization plateau can last extremely long



$$\delta_m = \kappa = \log \frac{1-\epsilon}{\epsilon}$$

On the leading-order plateau dynamics:

$$T = \Theta\left(\frac{1}{\alpha_0 \log(1/\alpha_0)}\right).$$

- Small initial retention makes escape slow.
- $\alpha_0 = e^{-\Delta e^{\lambda_0}}$ creates severe distance sensitivity.
- The theorem concerns reduced dynamics, not full Mamba training.

Paper, Fig. 1a–b, PDF p. 5; Proposition 4 / Eq. (B.4), PDF pp. 17–18.

In long contexts, retained keys also contribute noise

Now use fresh Gaussian keys and independent random labels.
Population gradient flow from zero gives $w_x = 0,\ v > 0$.

$$g = \underbrace{\alpha_\Delta \|x_q\|^2 \widetilde{y}_q}_{\text{signal}} + \underbrace{\sum_{t \neq t^*} \alpha_{\ell-t+1} (x_q^\top x_t) \widetilde{y}_t}_{\text{interference}} .$$

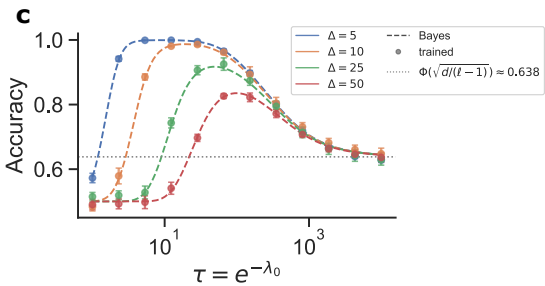
With $S_2 = \sum_{j=1}^{\ell} \alpha_j^2$, the conditional error is

$$\Pr(\operatorname{sign}(g) \neq \widetilde{y}_q \mid x_q) = \Phi\left(-\alpha_\Delta \|x_q\| \sqrt{\frac{d}{S_2 - \alpha_\Delta^2}} \right) .$$

Weaker decay increases both the retained signal and the noise variance.

Paper, App. B.2, Proposition 5 and Lemma 6, PDF pp. 18–19. Bayes error is for statistic g.

No decay is not always the best retrieval regime



As $\tau \to \infty$ at fixed ℓ:

$$p_{\tau,\Delta}(x_q) \to \Phi\left(-\|x_q\|\sqrt{\frac{d}{\ell-1}}\right).$$

For $d = 128, \ell = 1024$,

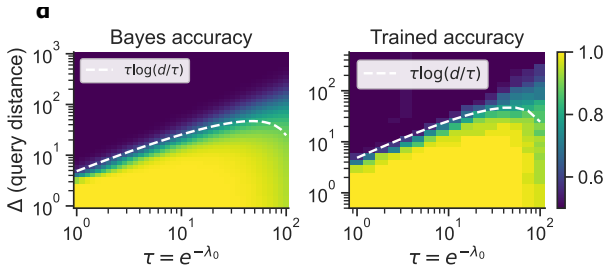
accuracy ≈ 0.638.

Error approaches $1/2$ only as $(\ell-1)/d \to \infty$.

The formal appendix uses $d/(\ell-1)$; the main-text formula has a typo.

Paper, Fig. 1c, PDF p. 5; formal Proposition 8, PDF p. 20; compare Sec. 2.4, PDF p. 4.

Decay supports local retrieval inside a much longer sequence



$$S_2 = \Theta(\tau), \qquad \mathrm{SNR} \asymp e^{-\Delta/\tau}\sqrt{d/\tau}, \qquad \Delta_{\max} = \Theta\left(\tau \log \frac{d}{\tau}\right).$$

- Assumptions: $\ell/\tau \to \infty$, $1 \le \tau = o(d)$, $\|x_q\| = \Theta(1)$.
- Arbitrarily many distant distractors can be tolerated if the target stays local.

Paper, Fig. 1d, PDF p. 5; Lemma 7 and Proposition 9, PDF pp. 19–21. Distances also cannot exceed ℓ.

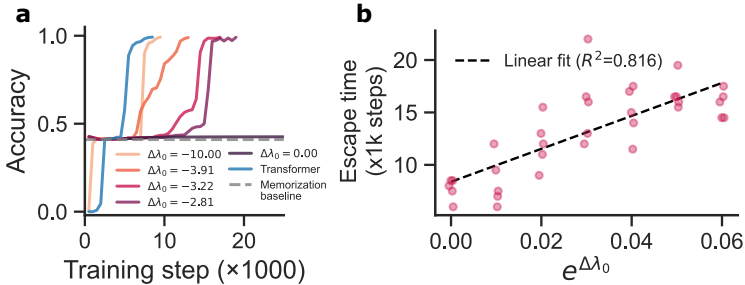
Three experiments test progressively less idealized retrieval

Task	Shortcut / intervention	What the comparison tests
Multi-token KV	Fixed values with token perturbations; gate offset	Escape from memorization; early vs. late targets
Logical rules	Fixed clauses, changing numeric names	Retrieve a current name; extrapolate with local relevant clauses
Tool calling	Pretrained Mamba; decay-rate scale before SFT	Correct calls and functions hallucinated outside the pool

Keep target position, pool size, and training distribution separate.

Paper, Secs. 3–5, PDF pp. 6–10. Each experiment uses different architectures and optimization.

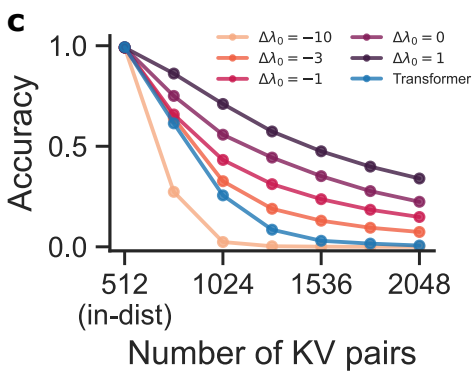
Weak gate initialization escapes the KV shortcut faster



- Eight-token values, flip probability 0.1: memory baseline $0.9^8 \approx 0.43$.
- Default initialization does not escape within 100k steps.
- Five seeds: panel a is the median-escape run; b shows all seeds.

Paper, Fig. 2a–b, PDF p. 7; App. D.2, PDF pp. 23–24. Escape means sequence accuracy > 0.95.

With later targets, stronger gates degrade more slowly

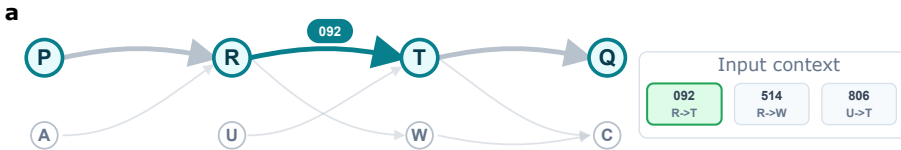


- Train with at most 512 KV pairs; test up to 2048.
- Query the last 10% of keys (best-last).
- Stronger gates retain higher accuracy as the pool grows.
- Five-seed means; appendix specifies 95% confidence intervals.

Better extrapolation still leaves substantial errors.

Paper, Fig. 2c, PDF p. 7; App. D.2.3, PDF p. 24. This is a separate best-last task.

Rule names separate context use from a memorized association

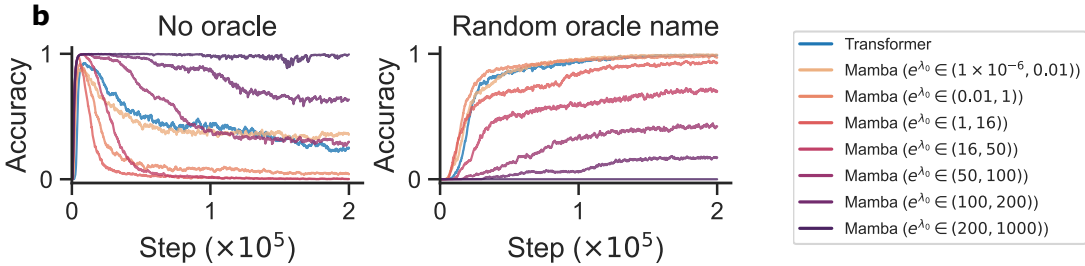


- Select a named clause that moves toward the goal.
- Actual task: 3 proposition sets, 128 propositions per set, 2 turns.
- **No oracle:** do not inject the correct clause; probes memorization.
- **Random oracle name:** inject a correct clause with a new name; probes retrieval.

Score the first turn as correct if the selected transition can still reach the goal.

Paper, Fig. 3a, PDF p. 8 (conceptual diagram); Sec. 4.2, p. 9; actual task in App. D.3.3, p. 25.

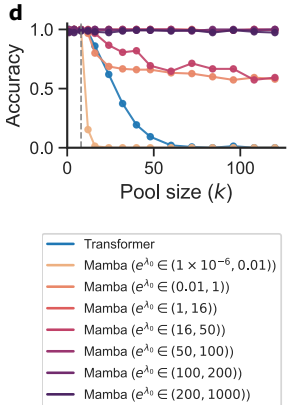
Strong gates preserve old names longer during training



- Strong decay: high no-oracle accuracy, slow learning of renamed clauses.
- Weak decay: faster adoption of the name supplied in the context.
- Train with pools up to 32; rename the first 10% of clauses.

Paper, Fig. 3b and legend, PDF p. 8; Sec. 4.2, p. 9. This is a diagnostic pair, not an exact decomposition.

Strong gates extrapolate when useful rules stay local



- A separate experiment: all names are resampled.
- Training pool size ≤ 8; evaluate up to 128 clauses.
- Rank useful clauses closer to the query.
- Strong gates stay near perfect accuracy; weak gates and this Transformer degrade.

This tests local retrieval amid more distractors, not arbitrary distant recall.

Paper, Fig. 3d and original legend, PDF p. 8; Sec. 4.2, p. 9; App. D.3, pp. 24–25.

BFCL tests a gate intervention after pretraining

- Mamba-2 1.5B vs. Llama-style Transformer 2B, each pretrained on 300B tokens.
- SFT on Nemotron Agentic tool calling; train pools up to $k = 24$.
- Modified BFCL `python simple`: BM25 pool, best-first order; inject the correct tool if missed.
- Before SFT, scale Mamba's decay rate:

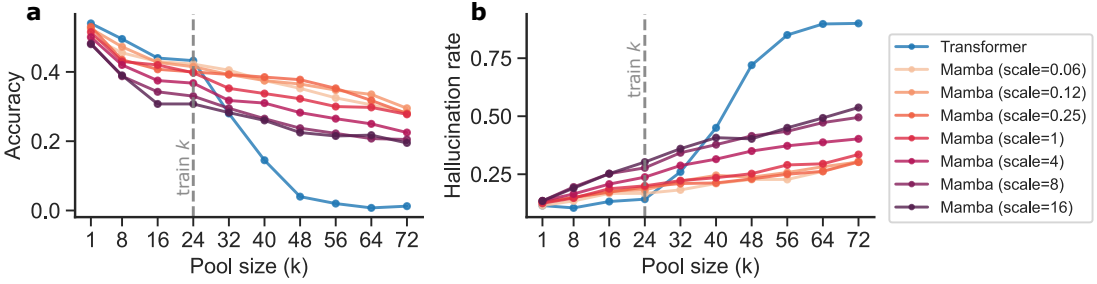
$$e^{\lambda} \leftarrow s e^{\lambda} \quad \Longleftrightarrow \quad \lambda \leftarrow \lambda + \log s.$$

Two distinct metrics

Accuracy: a fully correct call, including arguments.
Hallucination: a syntactically valid call to a function absent from the pool.

Paper, Sec. 5, PDF pp. 9–10; App. D.4, p. 25. Unequal model sizes; not the full BFCL benchmark.

Weaker Mamba gates improve calls and reduce pool violations

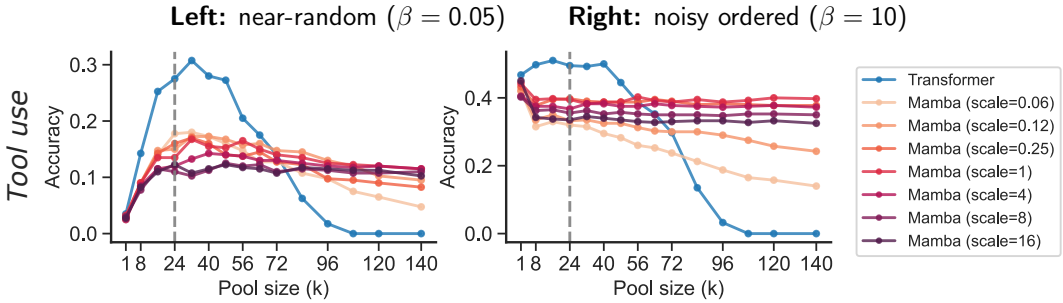


- Low-scale Mamba variants generally score better and hallucinate less.
- Some weakened gates improve on the pretrained scale $s = 1$.
- All Mamba variants extrapolate better than this Transformer here.

The authors do not find a clear gate-strength effect on length generalization in this range.

Paper, Fig. 4 and Sec. 5.2, PDF p. 10. No confidence intervals reported; small differences are not significance claims.

With imperfect tool retrieval, a larger pool has two effects



- Remove the guarantee that the correct tool is in the pool.
- Larger pools improve recall but can reduce conditional call accuracy.
- In the ordered setting, moderate gating remains useful as the pool grows.

Paper, App. C / Fig. 5 bottom, PDF p. 22. Left: near-random retrieval, $\beta = 0.05$; right: noisy ordered retrieval, $\beta = 10$.

The strongest evidence is mechanistic and task dependent

1. **Theory:** orthogonal or Gaussian keys, simplified readout, and population gradient flow; escape time uses reduced dynamics.
2. **Locality:** long sequences can remain manageable when targets stay close; arbitrary-position retrieval is a different requirement.
3. **Experiments:** controlled synthetic tasks and one modified BFCL split; model sizes are not always matched.
4. **Reporting:** five seeds for KV; fewer statistical details for the other studies and incomplete SFT hyperparameters.

Paper, Sec. 6, PDF pp. 10–11; Apps. B–D, pp. 15–25. Scope assessment by this reading.

Gate design must address learning and interference together

What this paper establishes

A retrieval solution can exist while optimization strongly favors a memorization shortcut.

Why removing decay is insufficient

Long contexts introduce interference; reducing it can require forgetting, with a cost to distant signals.

A useful next experiment

Jointly vary target distance, distractor count, gate schedule, and tool-pool ordering under the same training budget.

Paper, Secs. 2 and 6, PDF pp. 2–5 and 10–11. The proposed experiment is reader analysis, not a paper result.

Source and reading map

William L. Tong et al.
On the Importance of Gating: Memorization vs. In-Context Learning in State Space Models.
arXiv:2609.16540v1, submitted 15 September 2026.

arxiv.org/abs/2609.16540v1

- Mechanism and formal proofs: Sec. 2 and App. B.
- Experiments and protocols: Secs. 3–5 and App. D.
- Imperfect external retrieval: App. C.
- Companion files: Chinese reading notes, original paper, and figure provenance.

All empirical plots are author figures, with source page and panel identified. No new training experiments were run.

Backup: the escape-time proof uses the early plateau

On $\delta_m = \kappa$, the leading-order equations are

$$\dot{v} = \alpha\phi(\alpha v), \qquad \dot{\alpha} = (\alpha \log \alpha)^2 v \phi(\alpha v),$$

where ϕ is bounded above and below by positive constants before escape.

$$b = \log(1/\alpha), \quad \frac{db}{dv} = -b^2 v, \quad \frac{1}{b(v)} = \frac{1}{b_0} + \frac{v^2}{2}.$$

Since $v_T = \Theta(1)$ at $\alpha(v_T)v_T = \kappa$,

$$T = \Theta\left[\int_0^{v_T} \exp\left(\frac{b_0}{1 + b_0 v^2/2}\right) \mathrm{d}v\right] = \Theta(e^{b_0}/b_0).$$

The proof does not supply a uniform approximation theorem for full-model training.

Paper, Lemmas 2–3 and Proposition 4, PDF pp. 16–18.

Backup: experimental configurations must stay separate

Setting	Architecture / data	Optimizer / measurement
Theory	Identity projections; Fig. 1a: $d = 512$, $P = 32$, $\ell = 8$, $\Delta = 4$	SGD, batch 128, lr 0.1; Fig. 1b uses lr 0.3
KV retrieval	Mamba-2: 24 layers, width 768; 8-token keys and values; 10k pairs	AdamW, peak lr 3×10^{-4}; five seeds; escape at accuracy > 0.95
Logical rules	Mamba-2: 4 layers, 16 heads, head/state size 256	Muon, peak lr 10^{-3}, batch 128; first-turn accuracy
Tool calling	Mamba-2: 32 layers, width 2048; 1.5B vs. 2B Transformer	AdamW for pretraining and SFT; SFT details incompletely reported

Paper, App. D, PDF pp. 23–25. See companion notes for remaining settings and baseline architectures.

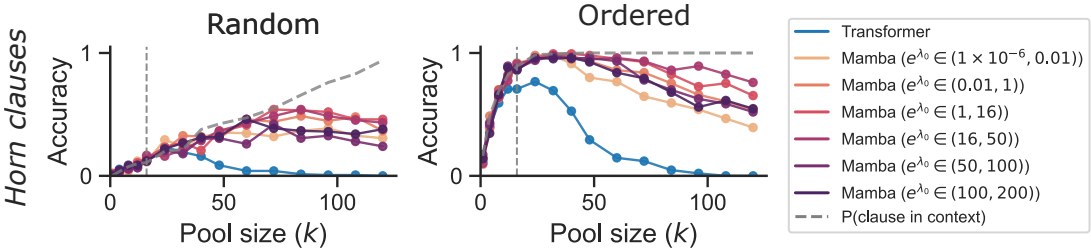
Backup: trained gates preserve the broad initialization ordering

Initial τ	After 200k	$\lvert\Delta\alpha\rvert$
2.1×10^4	1.2×10^4	4×10^{-5}
297	315	0.001
17	14	0.024
2.8	3.0	0.083
1.4	1.4	0.067
0.8	0.7	0.053
0.2	0.2	0.007

- Logical-rules length experiment.
- τ is the median effective length across heads, in tokens.
- Ordering survives a sweep of roughly five orders of magnitude.
- This supports a regime proxy, not an exactly static gate.

Paper, App. B.3 / Table 1, PDF p. 21. Values transcribed exactly; no recomputation from aggregate medians.

Backup: clause recall can rise while model accuracy saturates



- Left: random candidate ranking ($\beta = 0$); right: noisy ordered ranking ($\beta = 1$).
- Gray dashed curve: probability that a relevant clause enters the context.
- Recall gains alone do not guarantee end-to-end gains.

Paper, App. C / Fig. 5 top, PDF p. 22. The β scale differs from the BFCL experiment.

Backup: distinguish exact results from approximations

1. **Initial crossover.** With $s_0 = \sqrt{(1-2\epsilon)/P}$, equality of initial rates gives $\alpha = s_0/(1-s_0)$ when feasible. The paper uses the small-α approximation $\alpha^* \approx s_0$.

2. **No-decay limit.** Use $d/(\ell-1)$ from formal Proposition 8, not $d/\ell - 1$ printed in the main text.

3. **Offset scaling.** At fixed distance and base rate,

$$\alpha_\Delta(\delta) = \exp[-\Delta e^{\lambda_{\text{base}}} e^{\delta}].$$

   The linear fit of escape time against e^δ ($R^2 = 0.816$) is empirical; $\alpha \propto e^{-\delta}$ does not follow.

Reader algebra checked against Eq. (B.1), PDF p. 15; Sec. 2.4, p. 4; Proposition 8, p. 20; Sec. 3.2, p. 6.