Memory Mosaics at scale

Associative memory for learning from new context

Paper authors: Jianyu Zhang and Léon Bottou

Research group reading • 19 October 2026

arXiv:2507.03285v3 (14 January 2026) NeurIPS 2025

Similar standard benchmark scores, stronger learning from new context.

Three questions separate knowledge from learning

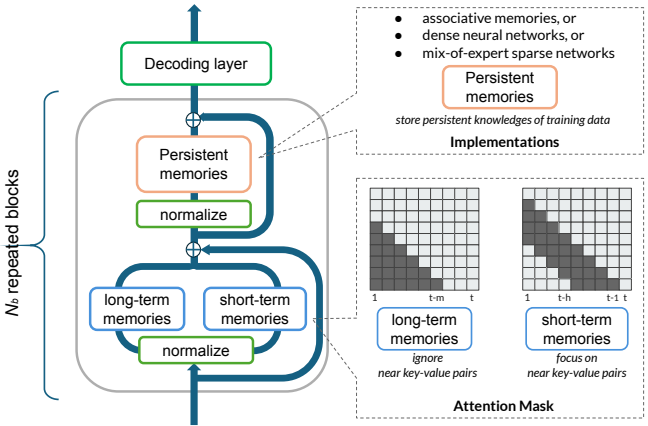
1. **Persistent knowledge:** What did pretraining teach the model?
2. **New knowledge:** Can it retain documents before seeing a question?
3. **In-context learning:** Can it infer a new input–label mapping from demonstrations?

Research question

Can associative memories improve the last two abilities at roughly 10B parameters, while preserving the first?

Paper, Secs. 1 and 5, PDF pp. 1, 5–7.

A block combines contextual and persistent memory



- Separate short- and long-term KV retrieval.
- Concatenate heads, then project with W_o.
- Persistent knowledge uses a SwiGLU feedforward network.

Paper, Fig. 2 (right) and Sec. 3.3, PDF p. 4. The experiments use dense SwiGLU persistent memory.

Retrieval estimates a conditional expectation

For stored pairs $\mathcal{M} = \{(k_i, v_i)\}_{i=1}^{n}$:

$$f(k; \mathcal{M}) \approx \mathbb{E}[V \mid K = k]$$

$$f(k; \mathcal{M}) = \sum_{i=1}^{n} \underbrace{\frac{e^{\beta k^\top k_i}}{\sum_j e^{\beta k^\top k_j}}}_{\text{similarity weight}} v_i.$$

- Keys are L_2 normalized; β controls retrieval sharpness.
- The current key also serves as the query.
- Key construction and causal masks still encode sequence order.

Paper, Eqs. (2)–(4), PDF p. 2. Gaussian distance gives a factor 2, absorbed into the dot-product β.

A value stores what followed a past situation

$$k_t = \varphi_\theta(x_t, x_{t-1}, \ldots), \qquad v_t = \alpha_\psi \,\mathrm{Norm}(\gamma W_\psi x_t + (1-\gamma) W_\psi x_{t+1}) \,.$$

Teaching example: $[A, B, \ldots, A]$

The earlier A provides a matching key; its value contains information about the following B.

- At position T, read only values indexed by $i < T$.
- Each retrieved v_i uses at most x_{i+1}, already observed by T.
- Complete v_T only after the next input arrives.

Paper, Eqs. (5), (9)–(10), PDF pp. 3–4. The causal timeline is an explanatory derivation.

Content gates decide what contributes to a key

$$\tilde{k}_t = W_\varphi x_t, \qquad g_t = e^{W_g x_t}, \qquad \lambda_t = e^{-|W_\lambda x_t|},$$
$$\bar{k}_t = g_t \tilde{k}_t + \lambda_t \bar{k}_{t-1}, \qquad k_t = \mathrm{Norm}\left(\bar{k}_t\right).$$

- $g_t > 0$ controls the current write; $0 < \lambda_t \leq 1$ controls retention.
- Both gates are scalars driven by the current content.
- Semantically irrelevant insertions can receive less weight.

Where the recurrence stops

The recurrence creates keys. The associative memory still stores historical KV pairs.

Paper, Sec. 3.2, Eq. (8) and footnote 6, PDF p. 3.

Larger memories use narrower retrieval kernels

$$\boxed{\beta(n) = \beta_1 n^{\alpha} + \beta_0}, \qquad \beta_0 \geq 0,\ \beta_1 > 0,\ 0 < \alpha < 1.$$

- n: number of available key–value pairs.
- More examples $\Rightarrow$ larger $\beta \Rightarrow$ smaller bandwidth $1/\sqrt{\beta}$.

Teaching calculation: one relevant key, $n-1$ distractors

If the similarity gap is $\Delta > 0$, the relevant weight is

$$a_{\star} = \frac{1}{1 + (n-1)e^{-\beta\Delta}}.$$

Increasing sharpness can counter dilution as memory grows.

Paper, Sec. 3.1, Eq. (6), PDF p. 3. Calculation illustrates the direction; it does not prove optimality.

Overlapping time scales support different roles

$$\mathcal{M}_t^S = \{(k_i^S, v_i^S) : t - h + 1 \le i \le t - 1\},$$
$$\mathcal{M}_t^L = \{(k_i^L, v_i^L) : 1 \le i \le t - m\}.$$

- **Short term:** local relationships within window parameter $h = 256$.
- **Long term:** distant context, excluding the nearest positions.
- Independent parameters; $m < h$ creates overlap.
- Train with random $m \in [64, 256]$; infer with $m = 64$.

Persistent memory stores learned parameters; long-term memory stores context KV pairs.

Paper, Secs. 3.3–4, Eq. (10), PDF pp. 4–5; Appendix C, PDF p. 14.

The main comparison matches training recipes

Configuration	small	large
Layers / hidden width / heads	24 / 2048 / 16	32 / 4096 / 32
Pretraining tokens	200B	1T
Pretraining $\rightarrow$ adapted context	4K $\rightarrow$ 32K	4K $\rightarrow$ 32K

- Baseline: Llama-style Transformer with multi-head attention.
- Same data mix and hyperparameters; document-wise training masks.
- Large parameter counts: **8.8B Transformer vs. 9.9B MM v2**.
- At length 256: 16.7B vs. 18.9B FLOPs/token.

Paper, Sec. 4 and Table 2, PDF pp. 4–5; Appendix C, PDF p. 14. FLOPs are short-context estimates.

Standard benchmark averages are nearly unchanged

19-benchmark average	Transformer	MM v2
small	37.8	38.0
large	52.2	52.2

Ablating long-term memory separates two groups

	Full MM v2	Without long-term memory
13 benchmark average	56.8	56.6
Other 6 benchmark average	42.1	34.9

The first group is a proxy for persistent-knowledge ability. The second group needs more contextual support.

Paper, Tables 1–2, PDF p. 5; Table 9, PDF p. 16. Scores in percent; large model for the ablation.

At 32K, multi-document QA improves by 12–15 points

Read multiple unrelated documents, then answer a question about one of them.

32K task length	Transformer	MM v2	Difference
small	22.1	36.9	+14.8 pp
large	41.1	53.4	+12.3 pp

- Both architectures were adapted from 4K to 32K context.
- The question arrives after the information must be stored.
- Simple single-needle retrieval nearly saturates: 99.6 vs. 100.0 for large models at 32K.

Paper, Sec. 5.2 and Table 4, PDF p. 6; Table 13, PDF p. 17. QA scores in percent; pp = percentage points.

Length extrapolation works, with declining accuracy

4K pretrained model	4K test	8K	16K	32K
Transformer large	57.7	×	×	×
MM v2 large	59.3	48.8	46.4	26.5
MM v2 small	45.0	35.0	34.1	31.7

- No length adaptation for the rows above.
- After 32K adaptation, MM v2 large reaches **46.4 at 64K**.
- × preserves the paper's failure marker; it is not a numerical zero.

Interpretation

Extrapolation retains useful retrieval, but longer contexts remain harder.

Paper, Tables 3–4, PDF p. 6. QA scores in percent.

Ablations connect long-term memory to retrieval

Remove the long-term branch

Large model, adapted to 32K:

$$53.4 \longrightarrow 20.2$$

32K QA: -33.2 pp

Randomize the training boundary

Small model, trained at 4K:

$$15.9 \longrightarrow 31.7$$

32K QA: $+15.8$ pp

The second comparison uses no length fine-tuning. Random overlap helps avoid dependence on a fixed distance boundary.

Paper, Tables 10 and 12, PDF p. 17. The two ablations use different scales and adaptation settings.

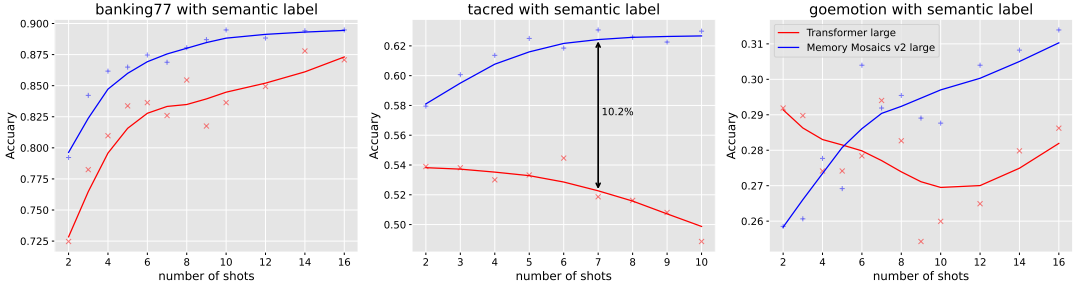
One shot means one example from every class

Task	Classes	Mean tokens/example
Banking77: intent	77	24
Tacred: entity relation	41	77
Goemotion: emotion	28	26

- Banking77, 16 shots: **1,232 demonstrations**.
- Semantic labels: use meaningful names such as `joy`.
- Anonymous labels: learn mappings to names such as `class_00`.
- Choose the best delimiter/order strategy for each *n*-shot setting.

Paper, Sec. 5.3 and footnote 13, PDF p. 7; prompt examples in Appendix I, PDF pp. 19–21.

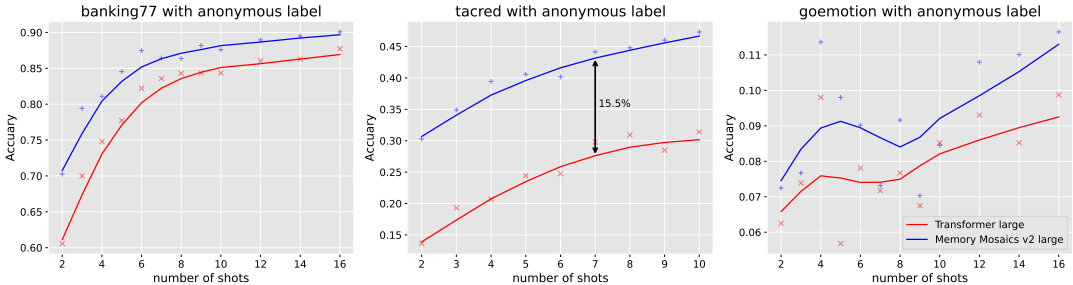
Semantic-label gains depend on the task and shot count



- Tacred benefits strongly; the marked gap is about **10.2 pp**.
- Banking77 improves for both models; low-shot Goemotion favors the Transformer.

Paper, Fig. 3, PDF p. 8. Large models; blue = MM v2, red = Transformer. Accuracy uses different vertical scales.

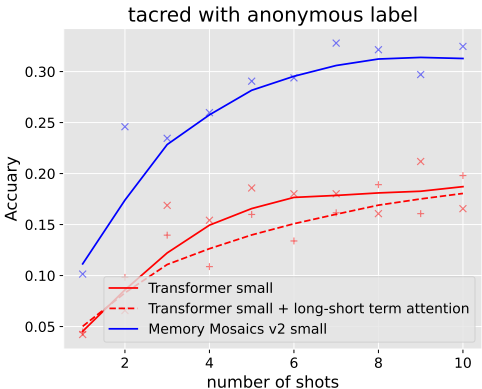
Anonymous labels reveal stronger mapping from examples



- MM v2 leads across the displayed curves; Tacred's marked gap is **15.5 pp**.
- Goemotion remains difficult: accuracy is roughly 0.07–0.11 for MM v2.

Paper, Fig. 4, PDF p. 8. Large models. The curves are not all monotonic in shot count.

Splitting Transformer attention alone is insufficient



- Long/short attention alone does not close the anonymous Tacred gap.
- This motivates studying the full combination of key, value, and retrieval design.
- A full ablation of all three v2 modifications is still missing.

Paper, Fig. 5, PDF p. 8. Small models. The last point is a reading assessment.

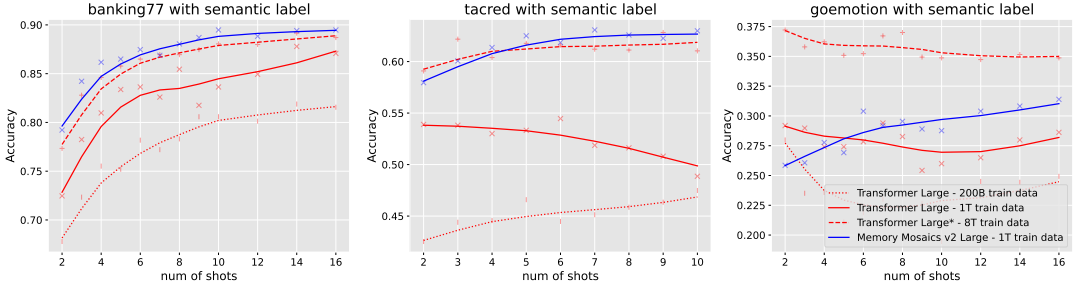
Eight times more data does not close the QA gap

Large model	Training tokens	32K QA
Transformer	200B	33.8
Transformer	1T	41.1
Transformer*	8T	46.9
MM v2	1T	**53.4**

- 8T Transformer still trails by 6.5 percentage points.
- Transformer* also uses GQA and 8K pretraining context.
- This comparison changes more than the training-token count.

Paper, Sec. 6.1 and Table 5, PDF p. 9. All rows use the 32K-adapted models.

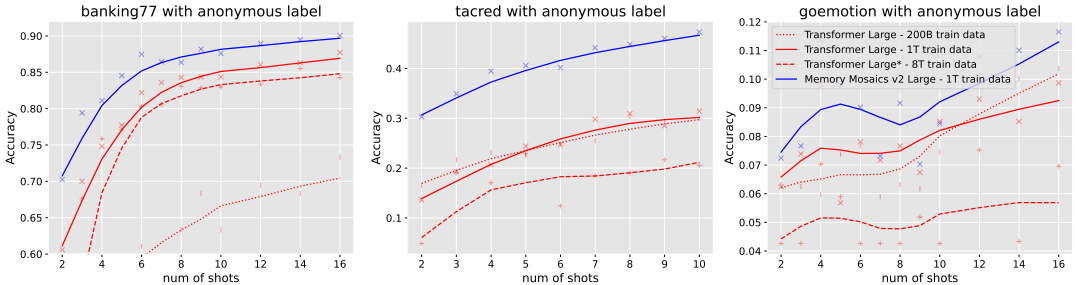
More pretraining can match semantic-label performance



- Dashed red: 8T Transformer. Blue: 1T MM v2.
- The 8T model reaches comparable results and leads on Goemotion.

Paper, Fig. 6, PDF p. 9. Meaningful class names allow use of pretrained semantic knowledge.

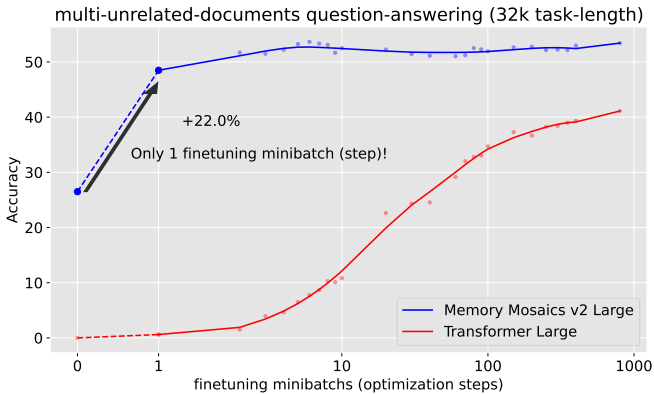
Anonymous-label performance remains harder to recover



- The 8T Transformer remains below MM v2 on all three curves.
- It also falls below the 1T Transformer in these displayed tasks.

Paper, Fig. 7, PDF p. 10. The 8T model also differs in attention type and pretraining context length.

One optimization step greatly improves length adaptation



- 4K $\rightarrow$ 32K adaptation.
- MM v2: about $+$**22 pp** after one mini-batch.
- Two steps approach the plateau.
- This is an optimization step, not one example or weight-free ICL.

Paper, Sec. 7 and Fig. 8, PDF p. 10. Transformer at 800 steps remains below MM v2 at one step.

The capability gain has a concrete storage cost

- **Capacity:** full contextual KV storage grows with context length.
- **Computation:** exact long-term retrieval still scans history; full-sequence interactions remain quadratic in length.
- **Budget evidence:** Appendix E compares parameter/FLOP curves, but gives no serving throughput or latency results.
- **Mechanism evidence:** long-term memory and random overlap have ablations; separate contributions of all new modules remain open.

Paper, footnote 6, PDF p. 3; Sec. 8, PDF p. 10; Appendix E, PDF p. 16. Complexity statement follows from exact KV retrieval.

The main contribution is better use of new context

1. Associative-memory design scales to **9.9B parameters and 1T tokens**.
2. Comparable standard benchmarks can hide large differences in **new-knowledge retrieval and ICL**.
3. More pretraining helps semantic tasks, but does not recover all **anonymous-label learning** in these comparisons.

A useful next experiment

Separate the effects of successor values, gated keys, and adaptive bandwidth under matched budgets; then reduce long-term retrieval cost.

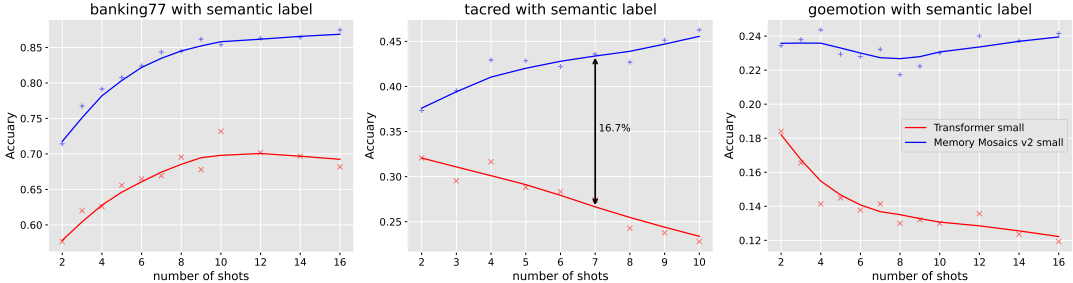
Jianyu Zhang and Léon Bottou. Memory Mosaics at scale. NeurIPS 2025.
arXiv:2507.03285v3, revised 14 January 2026. The proposed experiment is a reading perspective.

Backup: six benchmarks rely more on long context

Benchmark	Full MM v2	No long-term memory	Difference
SQuAD	78.2	69.4	−8.8
BBH	47.8	24.6	−23.2
MATH	8.8	5.4	−3.4
MBPP	9.6	6.8	−2.8
RACE middle	61.6	59.5	−2.1
RACE high	46.5	43.6	−2.9
Average	42.1	34.9	−7.2

Paper, Table 9, PDF p. 16. Large model; scores in percent and differences in percentage points.

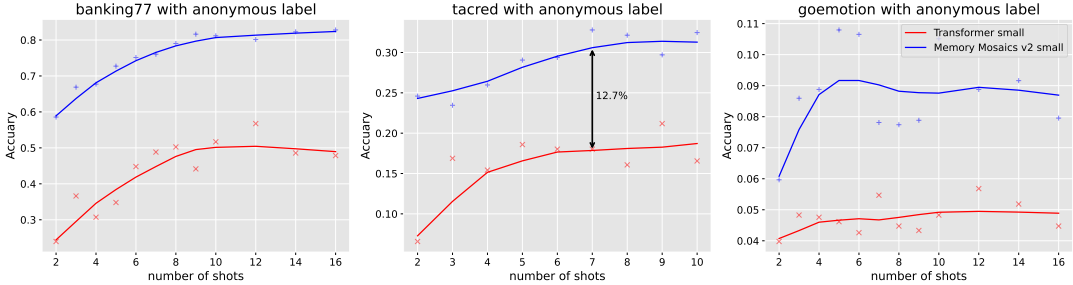
Backup: the small model also benefits on semantic ICL



- Tacred's marked gap is approximately 16.7 percentage points.
- Small-model Goemotion is relatively flat for MM v2 as shots increase.

Paper, Fig. 13, PDF p. 18. Small models trained on 200B tokens.

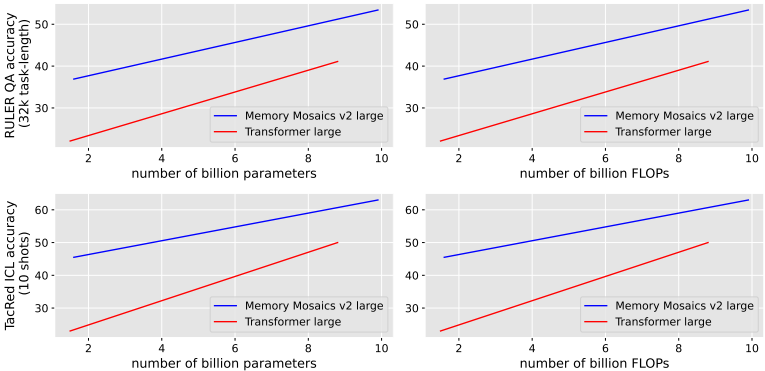
Backup: anonymous ICL gains persist at small scale



- Tacred's marked gap is approximately 12.7 percentage points.
- Anonymous Goemotion stays low in absolute accuracy for both models.

Paper, Fig. 14, PDF p. 18.

Backup: performance–budget curves favor MM v2



Sparse scale points connected by curves; intermediate-budget paired runs and the FLOPs counting convention are not detailed.

Paper, Fig. 12, PDF p. 16. Top: 32K RULER QA. Bottom: 10-shot Tacred. This is not a latency benchmark.

Backup: initialization makes the parameter constraints explicit

Parameter	Reparameterization	Initial θ
β_0, β_1	$e^{\min(\theta,10)}$	1.5
α	$\min(\lvert\theta\rvert, 1)$	$1/3$
α_ψ	$e^{\min(\lvert\theta\rvert,15)}$	0
γ	No reparameterization	$U(0,1)$ directly

- The implemented exponent can reach 0 or 1.
- γ starts in $(0,1)$, but no ongoing interval constraint is specified.
- Pretraining: AdamW, batch 1024, 2000 warmup steps, cosine decay.

Paper, Appendix C and Table 6, PDF p. 14. Initialization details refer to pretraining.

Backup: external baselines provide supporting context

- Appendix D reuses published RWKV and Mamba results with poor long-context retrieval and Tacred ICL.
- These are particular models and protocols, not a proof that all compressed-memory architectures must fail.
- At 32K QA, public-model scores include Mistral-base 44.7 and Mixtral-base 41.3, versus MM v2 53.4.
- Public comparisons differ in training and generation length: 128 output tokens for the other models, 32 for MM v2.

Paper, Tables 7–8, PDF p. 15; Table 11, PDF p. 17. Controlled in-paper Transformer comparisons remain the main evidence.