Deep Fusion Memory

A Chunk-Level Parametric Memory for Large Language Models

Paper by Rubin Wei, Siyuan Sun, Hao Sun, Yuliang Liu,
Xiang Shi, Jiarui Wang, Junming Zhang, and Zhouhan Lin

arXiv:2610.25100v1 Paper discussion, 26 October 2026

Publication not found as of 26 October 2026.
Code forthcoming, according to the paper.

The paper asks how to learn external knowledge without online search

Research question

Can retrieval supervise a compact, reusable memory that enters the decoder's hidden states, while the backbone stays frozen?

- Token-level parametric memory needs extensive **offline retrieval**.
- Output-probability interpolation gives memory no further decoder layers to work through.
- DFM predicts **one memory set per chunk** and uses gated cross-attention inside selected layers.

Paper, Secs. 1–3, PDF pp. 1–5.

DFM changes both supervision granularity and fusion location

Approach	Knowledge access	Decoder interface
CPT / LoRA	Train on corpus	Adapt model weights
RAG / RETRO	Retrieve at inference	Prompt or cross-attention
MLP Memory	Learn offline retrieval	Output distribution mix
DFM	**Predict chunk memory**	**Hidden-state fusion**

Key architectural distinction

A frozen backbone can process the injected memory through subsequent decoder layers.

Paper, Secs. 1–2, PDF pp. 1–3.

Three components turn offline neighbors into internal memory

1. Offline teacher

Retrieve ranked chunk embeddings from the training corpus. These become target memory V_j.

2. Causal predictor

Predict $\widehat{V}_j$ from preceding chunks; no retrieval is needed at inference.

3. Hidden-state fusion

Project predicted slots to M_j and read them through gated cross-attention.

$$X_{j-1} \xrightarrow{\text{retrieve}} V_j \quad \Big| \quad X_{<j} \xrightarrow{\text{predict}} \widehat{V}_j \xrightarrow{\text{project}} M_j$$

Paper, Sec. 3, Fig. 1, PDF pp. 3–5. See backup slide for the original schematic.

One retrieved target supervises a full 32-token chunk

1. Partition training text into chunks $X_0, \ldots, X_{J-1}$ of $C = 32$ tokens.
2. For $j \geq 1$, retrieve neighbors using the **previous** chunk X_{j-1}.
3. Store ranked Top-16 chunk embeddings as $V_j \in \mathbb{R}^{16 \times 1024}$.

X_0 has no previous-chunk query and receives no memory supervision.

Paper, Sec. 3.1 and Appendix A.1, PDF pp. 3–4, 14.

Offline teacher

Frozen Qwen3-Embedding-0.6B; normalized vectors; training-only FAISS datastore.

Why this saves queries

About $N/32$ queries for N training tokens, rather than one per token.

A causal predictor replaces retrieval at inference

$$\widehat{V}_j = \text{Predictor}_\phi(C_{\leq j}) \in \mathbb{R}^{K \times d_m}$$

- C_j mean-pools frozen token embeddings from X_{j-1}.
- Causal attention lets $\widehat{V}_j$ use only $X_0, \ldots, X_{j-1}$.
- K output heads predict the Top-K ranked memory slots.
- Default predictor: 28 layers, width 1,024, $K = 16$.

The learned object

Continuous memory vectors, not retrieved passages or next-token distributions.

Paper, Sec. 3.2, Eq. (1), PDF p. 4; Appendix A.1, p. 14.

Gated cross-attention injects memory into selected layers

$$M_j = P_\psi(\widehat{V}_j) \in \mathbb{R}^{K \times d}$$

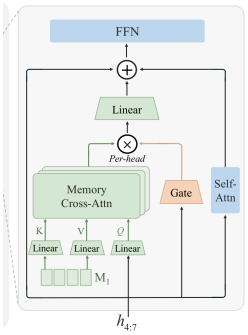
$$g_t = \sigma(W_g z_t + b_g)$$

$$u_t = W_O \,\mathrm{Concat}_{a=1}^{H}(g_{t,a} o_{t,a})$$

$$\widetilde{h}_t = h_t + \mathrm{SelfAttn}(z_{\le t}) + u_t$$

- Token state queries K shared slots; gate scales each head.
- Fusion layers: $[9, 16, 24, 31]$ (zero-based).

(c) Memory Fusion



Original Fig. 1(c); Paper, Sec. 3.2, Eqs. (2)–(3), PDF pp. 4–5.

The LM sees predicted memory throughout joint training

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \beta \mathcal{L}_{\mathrm{MSE}}, \qquad \mathcal{L}_{\mathrm{MSE}} = \frac{\sum_{j=1}^{J-1} \| \widehat{V}_j - V_j \|_F^2}{(J-1) K d_m}, \quad \beta = 1.$$

- Retrieval vectors V_j are **targets only**; they do not enter the LM forward pass.
- LM gradients flow through fusion back into the predictor.
- Train predictor, projector, cross-attention and gates; freeze the backbone and encoder.
- Gate weights start at zero; default bias -2 gives $\sigma(-2) \approx 0.119$.

Paper, Sec. 3.2, Eq. (4), PDF p. 5; Appendix A.1, p. 14.

Reported QA inference uses fixed prompt memory

General inference path

Input embeddings $\rightarrow$ memory predictor $\rightarrow$ selected decoder layers. No datastore lookup or external embedding encoder.

- In QA evaluation, the **visible prompt chunks** form memory queries.
- Predicted memory stays fixed while the answer is generated.
- Generated answer chunks do not refresh queries or receive direct memory updates.

The QA results therefore test prompt-conditioned memory during answer generation.

Paper, Sec. 3.3, PDF p. 5; Appendix A.1, PDF pp. 14–15.

The six-task QA mean combines different task metrics

Training and evaluation	Six benchmarks
2018 English Wikipedia; roughly 4B tokens; Mistral-7B-v0.3 and Qwen3-8B-Base.	NQ, WebQA; HotpotQA, 2Wiki, MuSiQue, Bamboogle.
Zero-shot greedy decoding; no online retrieval, gold documents or reasoning demonstrations.	F1 on five tasks; **substring match** on WebQA. The six scores receive equal weight.

Baselines: RAG, xRAG, RETRO, CPT, LoRA, MLP Memory.

Paper, Sec. 4.1 and Appendix A.1, PDF pp. 6, 14–15; Table 4.

DFM raises the six-task QA average on both main backbones

Backbone	Base	Best baseline	DFM	Gain over baseline
Mistral-7B-v0.3	13.63	xRAG 24.14	**27.21**	+3.07 points; +12.71% relative
Qwen3-8B-Base	22.71	MLP 23.80	**26.46**	+2.66 points; +11.19% relative

Read the percentages correctly

Relative percentage gain divides the *point difference* by the baseline mean; it is not a percentage-point change.

Paper, Sec. 4.2, Table 1, PDF pp. 6–7. Relative values as reported by the authors.

Task-level wins and failures explain the mean

Backbone	Task	Base	DFM	Relevant comparator
Mistral	Bamboogle	3.91	**22.65**	RAG 18.56
Mistral	2Wiki	5.87	25.00	**RAG 31.02**
Qwen3	2Wiki	24.83	**32.38**	MLP 27.47
Qwen3	WebQA	41.78	39.47	**MLP 43.26**

DFM leads in the **average**, not every task. Qwen3 WebQA falls below Base.

Paper, Sec. 4.2, Table 1, PDF p. 7. Bamboogle has 125 test examples (Table 4, p. 15).

A third QA backbone shows the same average trend

OLMo-3-1025-7B	Base	Best baseline	DFM
Six-task mean	17.99	RETRO 23.63	**25.49**

- DFM improves the mean by 7.50 points over Base and 1.86 over RETRO.
- It improves four of six tasks over Base.
- WebQA and HotpotQA decrease by 1.28 and 0.63 points.

Paper, Appendix B.1, Table 5, PDF p. 16.

Domain adaptation tests code, math and law separately

- Separate DFM checkpoint for each domain on Qwen3-4B-Base and Qwen3-8B-Base.
- Code and Math: 100M training tokens each; Law: 123.5M.
- HumanEval and HumanEval+ use execution correctness (estimated pass@1).
- GSM8K tests math; LawBench averages 20 legal task metrics.

Reported domain mean

Equal-weight average of HumanEval, HumanEval+, GSM8K and LawBench scores.

Paper, Sec. 4.1 and Appendix A.2, PDF pp. 6, 15.

DFM also has the highest domain mean at both scales

Backbone	Base	CPT	LoRA	DFM	DFM − Base
Qwen3-4B-Base	58.18	62.92	60.70	**64.69**	+6.51
Qwen3-8B-Base	63.14	65.06	66.96	**67.52**	+4.38

- 4B: CPT leads LawBench (41.66 vs DFM 40.41).
- 8B: LoRA leads HumanEval+ (65.85 vs 64.02) and GSM8K (87.19 vs 86.43).

Paper, Sec. 4.3, Table 2, PDF p. 8. All differences are points on the reported 0–100 scale.

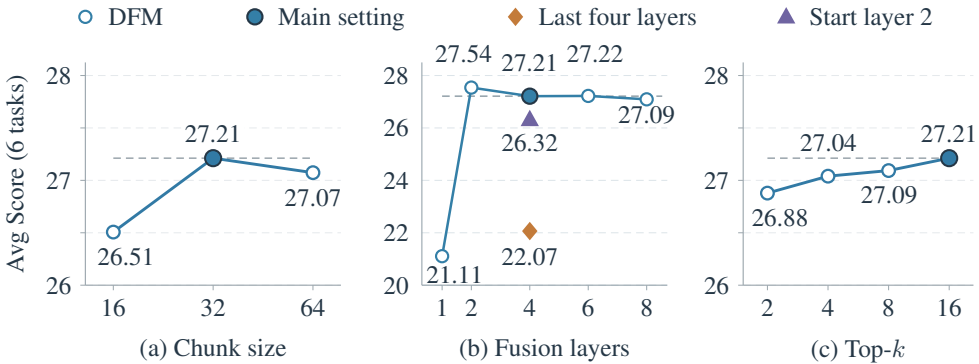
Construction counts improve far more than measured training time

Compared with MLP Memory	MLP Memory	DFM
Added parameters	1.50B	0.75B
Training on 16 A800 GPUs	64 h	24 h
Queries for 4B tokens	4B	125M
Returned neighbor slots	4.096T	2B

32× fewer queries; 2048× fewer returned slots because Top-1024 becomes Top-16 too.

Paper, Sec. 4.4 and Appendix D, Fig. 2, Table 13, PDF pp. 9, 23. Training hours exclude target construction and evaluation; slot counts are not timing measurements.

Fusion placement matters more than simply adding modules



- Four distributed layers: 27.21; final four layers: 22.07.
- Two layers at $[9, 31]$: 27.54; last layer alone: 21.11.

Original Fig. 3, PDF p. 10; Appendix C.1, pp. 20–21. Six-task means; single runs.

The default gate is conservative, but not the best sweep result

Gate configuration	Six-task mean	Difference vs default
No gate	26.83	-0.38
Bias $b_g = -2$ (default)	27.21	—
Bias $b_g = 0$	**27.90**	$+0.69$

- Zero-initialized gate weights and $b_g = -2$ start at $\sigma(-2) \approx 0.119$.
- Other tested variations are close: Top-2 memory 26.88; 14-layer predictor 27.31.

Paper, Sec. 5.2, Table 3, PDF pp. 10–11; Appendix C, pp. 20–22. Single runs.

The evidence supports average gains under a specific protocol

- QA means mix F1 with WebQA substring match; domain means mix four benchmarks.
- Offline retrieval **workload counts** do not establish end-to-end construction time or inference latency.
- QA answer generation uses fixed prompt memory, not refreshed answer-chunk memory.
- Ablations are single runs; Bamboogle has only 125 test questions.

The current corpus is learned into extra parameters; updating its facts would require further adaptation or another update mechanism.

Paper, Tables 1–4 and Appendices A.1, C, D, PDF pp. 7–15, 20–23. Final sentence is an architectural inference.

Takeaways

1. **One memory target per chunk** sharply reduces offline retrieval queries.
2. **Predicted continuous memory** can be fused into frozen decoder hidden states without inference-time search.
3. DFM improves reported **mean QA and domain scores**, with clear task-specific exceptions.
4. Fusion placement and gate initialization deserve attention; the paper does not time full inference or target construction.

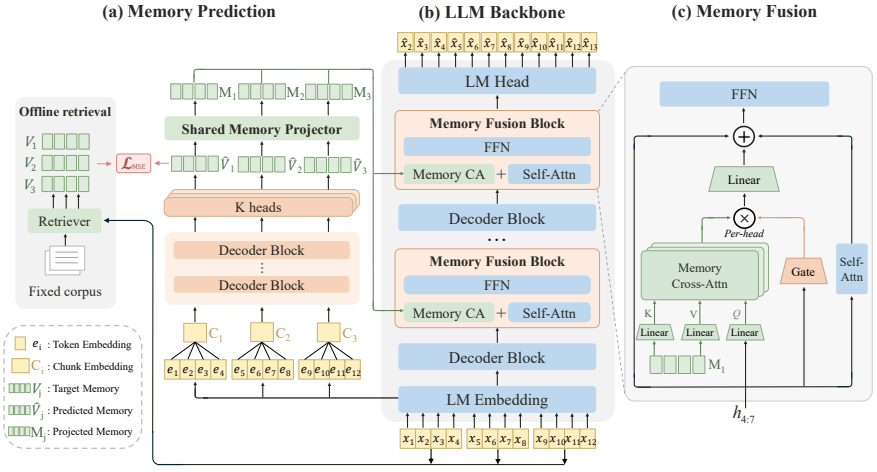
Paper, Secs. 3–6, PDF pp. 3–11; Appendix D, p. 23.

Paper and source links

- Rubin Wei, Siyuan Sun, Hao Sun, Yuliang Liu, Xiang Shi, Jiarui Wang, Junming Zhang, Zhouhan Lin. *Deep Fusion Memory: A Chunk-Level Parametric Memory for Large Language Models*. arXiv:2610.25100v1, 2026.
- Versioned paper PDF; author TeX source retrieved from the same version.
- Publication not found as of 26 October 2026; the paper says code will be released soon.

All figures in this talk come from the paper. Performance tables reproduce selected original values.

Backup: original architecture figure



Original paper figure, Fig. 1, PDF p. 4.