Language Models Need Sleep

Learning to Self-Modify and Consolidate Memories

Paper: Ali Behrouz, Farnoosh Hashemi,
Adel Javanmard, Vahab Mirrokni

arXiv:2606.03979v2 (10 July 2026)
Paper discussion prepared 29 September 2026

Periodically transfer fast memories into slower parameters,
then learn to generate useful adaptation data.

Independent reading presentation. 22 main slides + 5 backups; intended for 20–30 minutes.

The problem is retaining knowledge after context ends

Fast adaptation

New context changes model behavior.

Removing that context does not automatically preserve its contents in weights.

Persistent adaptation

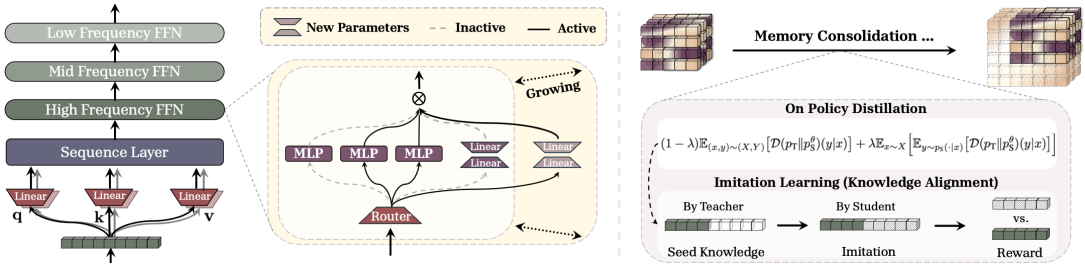
Weight updates can store new knowledge.

Repeated updates can interfere with earlier capabilities.

Research question: How can a model preserve a fast memory before rewriting it, while keeping capacity for future learning?

Paper, Secs. 1–2, PDF pp. 1–6. The sleep analogy motivates the design; it is not a biological equivalence claim.

Sleep separates consolidation from self-improvement



1. Consolidate

Expand a slower memory and distill into its new low-rank experts.

2. Dream

Generate, select and learn from synthetic adaptation data.

Original Fig. 2, PDF p. 7; Sec. 3.4, pp. 9–10. Figure depicts consolidation.

Different update intervals create a memory hierarchy

$$y_t = \mathrm{MLP}^{(f_k)}\Big(\cdots \mathrm{MLP}^{(f_1)}(x_t)\Big), \qquad C^{(\ell)} = \frac{\max_{\ell'} C^{(\ell')}}{f_\ell}.$$

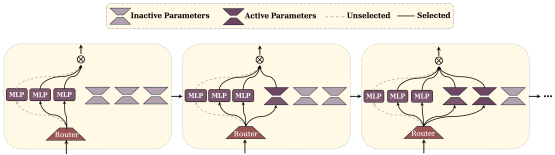
- f_ℓ: updates per reference time unit; $C^{(\ell)}$: interval between updates.
- Fast blocks adapt often; slow blocks retain information for longer.
- Before a scheduled fast-block update, consolidate its memory into the next slower block.

Keep frequency and interval distinct

The paper sometimes calls 1K/10K update intervals "frequencies." This ambiguity also affects the interpretation of Fig. 4.

Sec. 2.2, Definition 1 and Eqs. (1)–(2), PDF p. 5; Sec. 3.2, p. 7; Fig. 7, p. 22.

New experts provide room for incoming knowledge



Original Fig. 8, PDF p. 22.

$$A \in \mathbb{R}^{d \times d_{\text{low}}}, \qquad B \in \mathbb{R}^{d_{\text{low}} \times d}$$

- Add a low-rank expert in the receiving block.
- Freeze existing student weights during consolidation.
- Train only the new parameters.
- Preallocated, masked experts can implement growth.

Reading inference: frozen old weights do not guarantee unchanged old-task outputs; the forward computation still changes.

Secs. 3.2–3.3, PDF pp. 7–9. Matrix-pair cost is $2dd_{\text{low}}$, excluding routing and optimizer state.

Preserve the teacher until consolidation is complete

1. **Compute:** obtain prospective fast-block weights from accumulated gradients. Do not overwrite the running model.
2. **Consolidate:** teacher = pre-update model. Student = prospective fast weights, reset fast experts, and a new slow expert. Optimize the new expert to retain teacher behavior.
3. **Update:** apply the fast base-weight update, reset its old experts, and activate the trained slow expert.

Why this order matters

The teacher must still contain the memory that the student is about to lose.

Sec. 3.3, compute–consolidate–update protocol, PDF pp. 7–8. The teacher definition has a local layer-index typo.

Distill on both teacher and student trajectories

$$\mathcal{L}_{\mathrm{GKD}} = (1-\lambda)\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{F}(p_T \| p_S)(y \mid x)] + \lambda\mathbb{E}_{x\sim\mathcal{D},\, y\sim p_S(\cdot|x)}[\mathcal{F}(p_T \| p_S)(y \mid x)].$$

- $\mathcal{D}$: teacher-generated data; p_T, p_S: teacher and student.
- λ controls the share of student-generated trajectories.
- $\mathcal{F}$ compares output distributions along those trajectories.
- Stop gradients through sampling in this distillation step; update only expanded parameters.

Intuition: provide teacher feedback where the student actually goes.

Sec. 3.3, unnumbered GKD objective, PDF p. 8. Notation shortened; the paper leaves the divergence general.

Imitation rewards complement distribution matching

Learning to Imitate: sample a prefix of a teacher sequence, ask the student to continue, and reward agreement.

$$r = \gamma r_{\text{sem}} + (1 - \gamma) r_{\text{abs}}$$

- r_{sem}: binary semantic-equivalence reward from a frozen model.
- r_{abs}: thresholded, normalized Levenshtein similarity.

$$\underbrace{J_{\text{KS}}}_{\text{maximize}} = \mathbb{E}_{x \sim \mathcal{D},\, y \sim p_S} \left[(1 - \alpha) r(y) - \alpha \mathcal{D}(p_T \| p_S)(y \mid x) \right].$$

The paper names this expression $\mathcal{L}_{\text{KS}}$; its signs imply reward maximization, or minimization of its negative.

Sec. 3.3, Eqs. (3)–(4) and unnumbered KS objective, PDF p. 9. α is distinct from λ.

Dreaming learns which synthetic data improves adaptation

1. Generate candidate dreams from context C, with an extra random MoE expert during generation.
2. Select promising candidates using gradient information; retain random candidates for diversity.
3. LoRA-finetune an isolated model copy on each selected dream.
4. Reward the dream if downstream score τ improves; optimize generation using ReST^{EM}, following SEAL.

Reproducibility gap: a gradient vector is defined, but its conversion to a scalar ranking score is unspecified.

Scope: this loop requires an evaluation signal; self-generated data alone is insufficient.

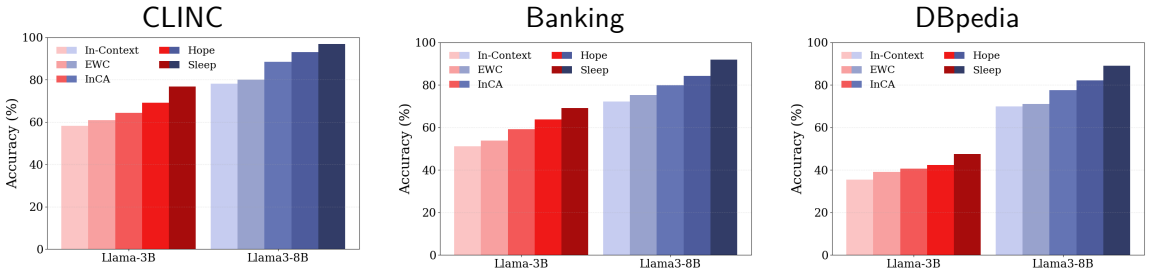
Sec. 3.4 and Eq. (5), PDF pp. 9–10. The dream utility reward differs from the imitation reward.

The experiments test several distinct capabilities

Question	Benchmark	Reported metric
Retain old classes?	CLINC / Banking / DBpedia	Accuracy ↑
Use long contexts?	MK-NIAH / LongHealth	Accuracy ↑
	QASPER	Perplexity ↓
Learn sequential skills?	Manchu / Kalamang	ChRF ↑
Scale context length?	BABILong	Accuracy ↑
Improve reasoning?	AIME / HMMT	average@16 ↑
Store new facts?	SQuAD, no passage at test	Accuracy ↑
Adapt from few examples?	Filtered ARC subset	Attempt success ↑

Sec. 4, PDF pp. 10–13; Appendix B, pp. 25–26. Early experiments emphasize consolidation; SQuAD and ARC use full Sleep.

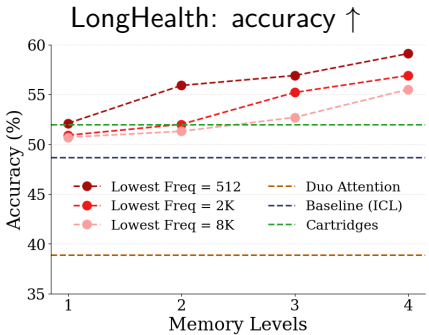
Sleep improves class-incremental accuracy over Hope

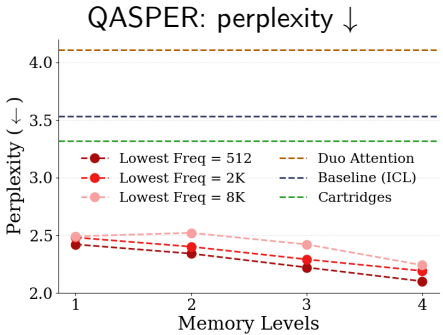


- Both 3B and 8B backbones show higher Sleep bars than Hope.
- ICL, EWC, InCA and Hope represent different retention mechanisms.
- End-point bars do not reveal a full forgetting trajectory; no error bars or exact per-bar numbers are supplied.

Original Fig. 3, PDF p. 10; Sec. 4.1, p. 11; dataset definitions, Appendix B.1, p. 25.

More levels generally help long-context behavior

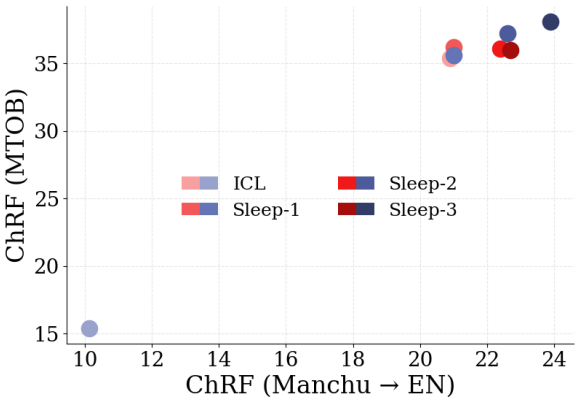




- Stronger performance with more memory levels is the overall trend.
- It is not strictly monotonic: QASPER's 8K curve worsens from one to two levels. Some LongHealth settings trail Cartridges.

Original Fig. 4 (middle/right), PDF p. 10. Frequency labels are reproduced as printed; their units are ambiguous.

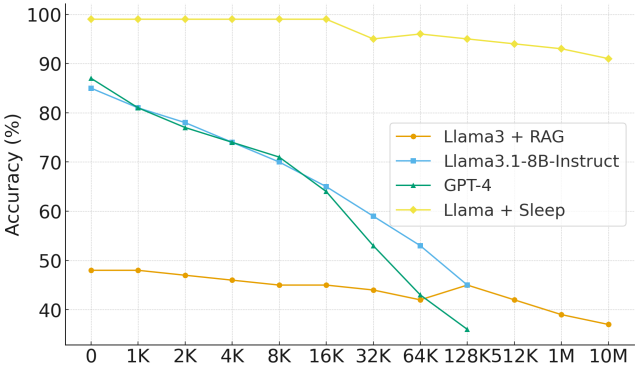
Sleep retains more after learning two new languages



- Both axes are ChRF; upper-right is better.
- Caption: red = single language; blue = sequential learning.
- Sequential ICL falls far below Sleep variants.
- Blue Sleep-2/3 points exceed their red points; the plot is stronger than "nearly recovering" alone.

Original Fig. 5, PDF p. 11; discussion, pp. 11–12. Two-language evidence, not an indefinite task stream.

BABILong remains strong at 10M, with qualifications



- Sleep is about **91%** at 10M (visual estimate).
- Small models receive benchmark-specific finetuning.
- The caption's red/blue explanation does not match the lines.
- Several baselines discussed in the appendix are absent from this figure.

Original Fig. 6, PDF p. 11; Sec. 4.1, p. 12; Appendix B.2, p. 26. Length bins are not plotted on a linear numeric scale.

Reasoning gains over OPSD are 1.0–2.6 points at 8B

Qwen3-8B; average@16 ↑

Method	AIME-24	AIME-25	HMMT-25
Base (Instruct)	73.8	68.1	42.4
SFT	75.5	66.4	43.7
GRPO	76.4	68.1	44.9
OPSD	76.6	67.4	45.1
Sleep	**79.2**	**69.0**	**46.1**
Sleep − OPSD	+2.6	+1.6	+1.0

- average@16 is not pass@16: do not interpret it as at least one success.
- The paper gives no uncertainty estimates for these differences.

Table 2, PDF p. 12. Values transcribed; differences calculated for this presentation. 1.7B results in backup.

Four-level Sleep improves no-context factual recall

SQuAD mean accuracy (%); original passage absent at evaluation

Method	One passage	CPT: 200 passages
Base	31.9	31.9
Finetuning, no Dreaming	33.4	32.0
SEAL	46.7	43.2
Sleep: two-level Transformer	48.1	44.3
Sleep: four levels	**48.9**	**46.2**

- Four levels vs. SEAL: **+2.2 / +3.0 percentage points**.
- CPT uses 974 questions; five dreams per passage form one training corpus.
- This is one 200-passage run, not 200 separately evaluated learning cycles.

Table 3, PDF p. 12; protocol, pp. 12–13. Backbone settings are referred to SEAL rather than fully restated.

ARC's 80% is an adaptation-attempt success rate

Method	Success (%)
ICL	0
TTT	10
SEAL	72.5
Sleep	**80**

- Llama-3.2-1B; filtered tasks: 11 train, 8 held out.
- Training: generate 60 dreams, reject 45 per task.
- Test: five dreams per task, each applied independently.
- Report the fraction of dream adaptations yielding a correct answer.

The denominator changes the interpretation

The protocol implies 40 attempts: 80% = 32/40, vs. 29/40 for SEAL. These counts are inferred, not individually reported.

Table 4, PDF p. 12; Sec. 4.1, p. 13; Appendix B.3, p. 26. Not a full ARC benchmark task pass rate.

Expansion explains part of the reasoning gain

Qwen3-8B; average@16 ↑

Method	AIME-24	AIME-25	HMMT-25
Sleep	**79.2**	69.0	**46.1**
Without imitation learning	76.8	67.9	45.0
Without semantic reward	78.9	**69.2**	44.5
Without expansion	78.2	67.9	44.9
OPSD	76.6	67.4	45.1
OPSD + expansion	77.9	68.2	45.9

- Sleep vs. expanded OPSD: only $+1.3$ / $+0.8$ / $+0.2$ points.
- Removing semantic reward slightly improves AIME-25; component gains are not universal across metrics.

Table 1, PDF p. 11; Sec. 4.2, p. 13 (its "Figure 1" reference should point to this table).

Dreaming has the largest SQuAD removal effect

Mean no-context SQuAD accuracy (%)

Configuration	One passage	CPT: 200 passages
Four-level Sleep	**48.9**	**46.2**
Remove gradient selection	47.1	45.2
Remove random expert	48.0	44.7
Remove Dreaming	35.7	36.2

- Relative to the four-level row: removing Dreaming costs **13.2 / 10.0 percentage points**.
- Selection and random experts show smaller gains.
- Removal rows do not restate the backbone; the four-level reference follows the table's ablation context.

Table 3, PDF p. 12; Sec. 4.2, p. 13. No repeated-seed or matched-compute uncertainty analysis is given.

Efficiency depends on what is held constant

Same number of steps

The authors report SFT as about **4**× **more efficient** than Sleep.

Matched target performance

Reported SFT wall-clock multipliers: **4.3**×**, 3.6**×**, 4.8**× on the three math benchmarks.

- These statements can coexist: a costly step may make more progress.
- Absolute times, hardware and exact target scores are not supplied here.
- Constant active parameter count does not imply constant total storage or training cost.

Appendix B.5, PDF p. 26; active-parameter claim, Appendix B, p. 25. These are author-reported efficiency ratios.

The evidence supports gains, not indefinite retention

- **Capacity:** experts grow or are unlocked; the slowest memory still has a finite budget.
- **Forgetting:** frozen weights and self-distillation do not prove invariance of earlier task behavior.
- **Feedback:** Dreaming needs a usable downstream evaluator, and imitation uses a reward model.
- **Reproduction:** gradient scalarization, schedules, routing and key objective hyperparameters are incompletely specified.
- **Evaluation:** filtered ARC tasks, BABILong finetuning and absent uncertainty estimates limit generalization.

Reading analysis based on Secs. 3–4 and Appendix B. No general zero-forgetting theorem is presented.

A useful design principle: consolidate before rewriting

1. Preserve a teacher containing the current memory; train a student that anticipates the next update.
2. Use slower, newly available parameters to carry that memory forward.
3. Separate retaining knowledge from learning which synthetic data improves future adaptation.

Best supported conclusion: the combined design improves several reported benchmarks under their specific protocols.

Next test (reading proposal): match total capacity and compute, then measure old-task retention across many Sleep cycles.

Ali Behrouz, Farnoosh Hashemi, Adel Javanmard and Vahab Mirrokni. Language Models Need Sleep: Learning to Self-Modify and Consolidate Memories. arXiv:2606.03979v2. Original figures credited to the authors, CC BY 4.0.

Backup: gains at 1.7B are small on AIME-25

Qwen3-1.7B; average@16 ↑

Method	AIME-24	AIME-25	HMMT-25
Base (Instruct)	49.8	34.5	25.7
SFT	47.3	36.1	22.9
GRPO	51.0	38.6	26.1
OPSD	51.6	40.0	28.1
Sleep	**53.2**	**40.2**	**29.3**
Sleep − OPSD	+1.6	+0.2	+1.2

The 0.2-point AIME-25 difference has no accompanying uncertainty estimate. It should not be called statistically significant.

Table 2, PDF p. 12. Differences calculated for this presentation.

Backup: the literal CMS update needs index care

$$\boldsymbol{\theta}_{i+1}^{(f_\ell)} = \boldsymbol{\theta}_{i}^{(f_\ell)} - \boldsymbol{e}_{i,\ell}$$

$$\boldsymbol{e}_{i,\ell} = \begin{cases} \displaystyle\sum_{t=i-C^{(\ell)}}^{i} \eta_t^{(\ell)} f(\boldsymbol{\theta}_t^{(f_\ell)}; x_t), & i \equiv 0 \pmod{C^{(\ell)}}, \\ 0, & \text{otherwise.} \end{cases}$$

- $f(\cdot)$ is an optimizer error component, such as a gradient; it is not the frequency f_ℓ.
- The printed sum has $C^{(\ell)} + 1$ terms if both endpoints are included.
- The intended accumulation interval needs clarification before a literal implementation.

Sec. 2.2, Eq. (2), PDF p. 5. Reproduced as printed, with an explicit indexing caveat.

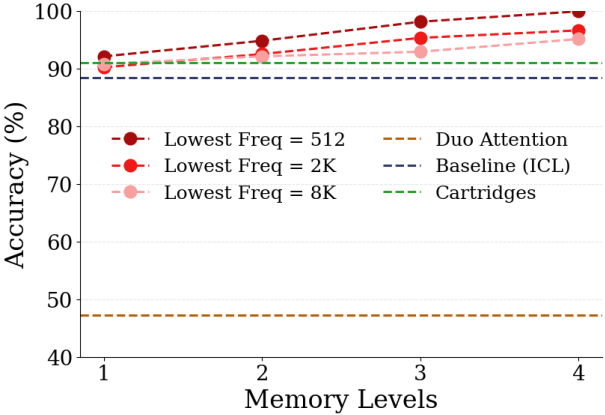
Backup: imitation uses two distinct similarity signals

$$r_{\text{abs}}(\hat{d}, d) = \begin{cases} 1 - \dfrac{z(\hat{d}, d)}{\max\{|\hat{d}|, |d|\}}, & z(\hat{d}, d) \leq z_0, \\ 0, & \text{otherwise.} \end{cases}$$

- z: Levenshtein distance; z_0: similarity threshold.
- r_{sem} is 1 for semantic equivalence, 0 otherwise.
- γ mixes these rewards; α mixes imitation and distillation.
- The paper does not fully report their task-specific settings or the semantic reward model's identity.

Sec. 3.3, Eqs. (3)–(4), PDF p. 9. This reward is distinct from Dreaming's post-adaptation improvement reward.

Backup: MK-NIAH improves with more memory levels



Original Fig. 4 (left), PDF p. 10; Sec. 4.1, p. 11; Appendix B.1, p. 25.

- Accuracy measures multi-key retrieval.
- Four-level, "512" curve reaches approximately 100%.
- ICL is around 88.5%; values here are visual estimates.
- The printed frequency labels lack a clear unit interpretation.

Backup: reported training settings and source scope

Setting	GRPO	SFT	Sleep
Learning rate	$5 \cdot 10^{-6}$	$5 \cdot 10^{-6}$	$5 \cdot 10^{-6}$
Effective batch	32	32	32
Training steps	500	100	100
LoRA rank	64	64	64
LoRA alpha	128	128	128

- Table 5 does not provide a separate OPSD column or a complete per-task configuration map.
- Read material: version-matched PDF, main TeX, all included body and appendix files, and original figures.
- This presentation analyzes reported evidence; no training results were independently reproduced.

Table 5, Appendix B.4, PDF p. 26. Bibliographic and source details are retained with the Chinese reading notes.