REMORY: Learning Residual Memory for Context Compaction

Research paper discussion | 20–30 minutes

Paper authors: Hanchen Xia, Baoyou Chen, Yutang Ge, Naihao Deng, Senqiao Yang, Zilong Dong, Weihao Yuan, and Siyu Zhu

arXiv:2610.11287v1 | Submitted October 8, 2026
Reading prepared October 13, 2026

Keep the summary; learn a continuous supplement for future decisions.

A readable checkpoint can lose future decision cues

The problem

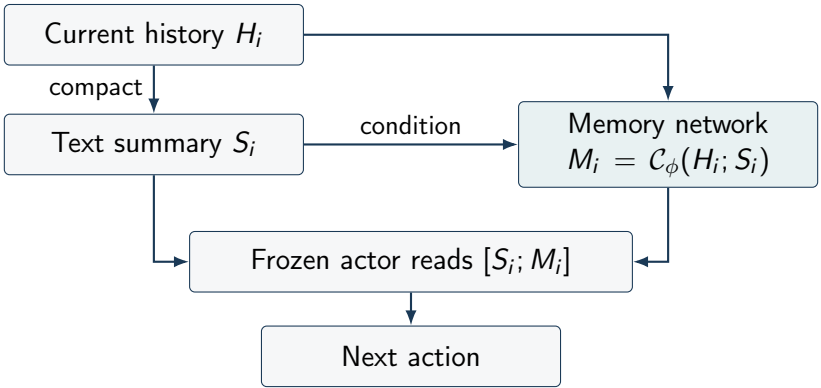
An agent must compact its history before it knows which details will matter later.

- A summary retains the main thread of a task.
- Source associations and earlier tool feedback can become useful later.
- Exact records can still be retrieved through ordinary tools.

Research question: can a learned, bounded supplement preserve useful access to history across compactions?

Paper, Sec. 1, PDF pp. 1–2.

The summary conditions the memory that follows it



Schematic redrawn from Fig. 3(a,b), PDF p. 3. Persistent instructions and ordinary archive tools omitted from the diagram.

Memory is a sequence of actor-input vectors

$$M_i = \mathcal{C}_\phi(H_i; S_i), \qquad |M_i| \leq m, \qquad m = 4{,}096.$$

- H_i: previous summary and memory, plus new interaction.
- S_i: the new textual summary at compaction boundary i.
- M_i: continuous vectors appended after the summary.

A recurrent state across compaction boundaries

The next history includes the new summary, new memory, and subsequent interaction. The same process repeats.

The residual is a predictive supplement along the sequence dimension.

Paper, Sec. 2.1, Eq. (1), PDF p. 3.

Block compression bounds the final memory

Up to 1,024 source positions $\longrightarrow$ 64 soft tokens

- Frozen actor features and learned queries predict vectors in parallel.
- Gated cross-attention makes construction summary-dependent.
- Hierarchical compression preserves source order and caps final slots.

Actor	Memory architecture	Trainable parameters
Qwen3.8-27B	Adapted DFlash	1.94B
GLM-5.3-Flash	Adapted DFlash-2	1.502B

Source encoding and intermediate representations can exceed the final cap.

Paper, Sec. 2.1, PDF p. 3. DFlash-2 includes grouped dynamic convolutions.

Stage I teaches the frozen actor to read soft memory

Teacher context	Student context
Source passage	Soft memory of the passage
+ instruction to repeat it	+ the same instruction

- Both predict the source text using the same frozen actor.
- Only the memory network learns to preserve source information.
- Data: text fields from MMFineReason-Full-2.3M.

Paper, Sec. 2.2, PDF p. 4.

Stage II learns a supplement for continuation

Teacher context	Student context
Original interaction history	Fixed summary + residual memory

- Both predict the same subsequent assistant response.
- The memory encoder never sees that future response.
- Initialize from Stage I; order examples by preceding-context length.

Reasoning and coding trajectories include Open-PerfectBlend prompts, precomputed summaries, and Qwen-generated responses.

Paper, Sec. 2.2, PDF p. 4.

Reverse KL trains the interface through a frozen actor

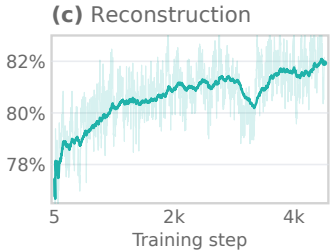
$$p_j = p_\theta(\cdot \mid C_{\text{full}}, y_{<j}),$$
$$q_j = p_\theta(\cdot \mid C_{\text{mem}}, y_{<j}),$$
$$\mathcal{L}(\phi) = \frac{\sum_j w_j D_{\mathrm{KL}}(q_j \| p_j)}{\sum_j w_j}.$$

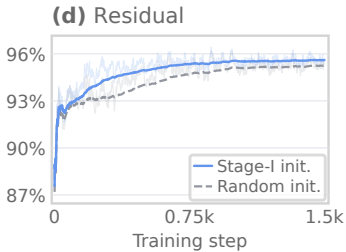
- Same target prefix $y_{<j}$; full-vocabulary distributions p_j, q_j.
- Positive w_j mildly emphasize early response positions.
- Update memory parameters ϕ; keep actor parameters θ fixed.

Summary content can overlap with memory: the objective rewards matching future predictions.

Paper, Sec. 2.2, Eqs. (2)–(3), PDF p. 4. All contexts include persistent instructions.

Reconstruction initialization helps residual optimization





- Axes: training step and teacher–student top-1 agreement.
- Right panel: same residual objective and data order, different initialization.
- These are training-stream diagnostics; the two panels use different tasks.

Paper, Fig. 3(c,d), PDF p. 3. Original vector panels, cropped with axes and legends retained.

Evaluation separates representation from agent behavior

Evaluation	Scale	What is controlled?
SummHay	92 queries, 10 collections	Same actor, query, summary, decoding; no retrieval
Agent tasks	4 benchmarks	Same actor, harness, tools, budgets, compaction policy
Next action	488 pairs from 100 trajectories	Same prompt and summary at the checkpoint

- Agent runs allow exact-history retrieval in both conditions.
- Context windows: Qwen 262,144; GLM one million positions.
- Agent intervention: add up to 4,096 residual slots at compaction.

Paper, Secs. 3.1–3.3, PDF pp. 4–6. Equal-position extra-text control is in SummHay.

SummHay gains are concentrated in source attribution

Coverage measures insight recovery; citation F1 measures source attribution.

Representation	Coverage	Citation F1	Joint score
Raw context	72.46	56.60	43.41
Summary	67.55	56.98	39.84
Summary++	67.52	56.94	39.78
Summary + REMORY	67.95	**61.03**	**43.39**

- Summary++ adds generated text under the same extra-position budget.
- Reported gains: coverage +0.40; citation F1 +4.04; joint +3.55 points.
- Joint gain: paired collection-bootstrap 95% interval $[+0.74, +6.21]$.

Table 1 and Sec. 3.1, PDF pp. 4–5. Higher is better. Gains use unrounded values; the interval is conditional on generated and judged outputs.

Near-full joint score with 5.2% of input positions

Representation	Mean input positions	Joint score
Raw context	96,872	43.41
Summary	1,393	39.84
Summary + REMORY	5,024	43.39

$$\frac{5{,}024}{96{,}872} \approx 5.19\%$$

- Relative to raw context: coverage is lower; citation F1 is higher.
- On 553 jointly covered insights, citation F1 rises 57.31 $\rightarrow$ 61.41.
- This pattern supports better links between recovered claims and evidence.

Paper, Sec. 3.1, PDF p. 5. The position count includes summary and memory; it is not a measured speedup.

Workflow gains persist across five runs

Qwen3.8-27B	Without memory	With memory	Gain
AutomationBench	35.5 ± 1.3	45.3 ± 1.0	+9.8
JobBench	33.4 ± 1.25	41.0 ± 1.41	+7.6

- AutomationBench: 600 tasks, scored by task verifiers.
- JobBench: 65 tasks, evaluated by matched GPT-5.6-Sol judges.
- Values are mean $\pm$ standard deviation over five runs.

Paper, Table 2 and Sec. 3.2, PDF p. 5. Scores are benchmark-specific; gains are score points.

A stronger actor with a larger window also benefits

Actor	Benchmark	Without	With	Gain
Qwen3.8-27B	Terminal-Bench 2.1	71.9	76.4	+4.5
Qwen3.8-27B	BrowseComp	74.0	77.0	+3.0
GLM-5.3-Flash	Terminal-Bench 2.1	84.3	87.6	+3.3
GLM-5.3-Flash	BrowseComp	84.9	89.0	+4.1

- Full task sets: 89 Terminal-Bench tasks; 1,266 BrowseComp tasks.
- One run per configuration for both actors.
- Published frontier references in Fig. 1 use different harnesses and protocols.

Paper, Table 3, PDF p. 5; Fig. 1, PDF p. 2. Scores are percentages; gains are percentage points.

Repeated outputs and tool errors both decrease

Actor	Benchmark	Fewer repeats	Fewer errors
Qwen	Terminal-Bench 2.1	17.9%	39.1%
Qwen	BrowseComp	29.8%	49.7%
GLM	Terminal-Bench 2.1	28.1%	25.3%
GLM	BrowseComp	29.8%	27.8%

- Repeats: identical outputs from the same tool within a task; planning and polling excluded.
- Errors: failed commands or explicit tool errors.
- These are aggregate counts; trajectory lengths and actions also differ.

Calculated from Table 3, PDF p. 5, as (without − with)/without. Definitions: Sec. 3.2, PDF pp. 5–6.

Generation spend usually falls; total cost is unmeasured

Actor / benchmark	Without	With	Change
Qwen / AutomationBench	\$0.59	\$0.53	−10.2%
Qwen / JobBench	\$3.09	\$3.21	+3.9%
Qwen / Terminal-Bench	\$4.29	\$2.86	−33.3%
Qwen / BrowseComp	\$5.39	\$4.69	−13.0%
GLM / Terminal-Bench	\$3.27	\$3.05	−6.7%
GLM / BrowseComp	\$5.65	\$5.00	−11.5%

Mean API-equivalent USD per task, using September 2026 uncached rates.
Residual positions count as actor input; source encoding, memory-network execution, and tools are excluded.

Paper, Tables 2–3, PDF p. 5; cost definition, PDF p. 6. Percentage changes calculated from reported costs.

With the checkpoint fixed, next actions are preferred

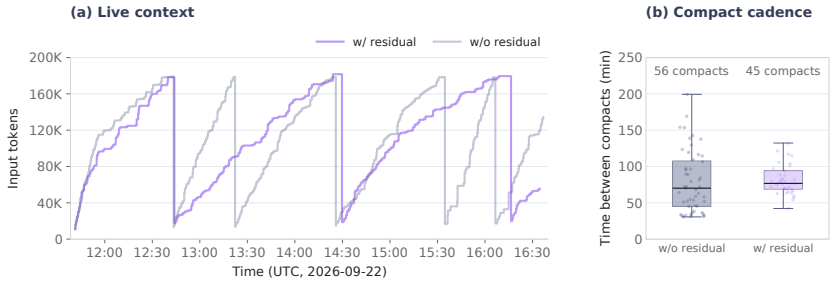
Residual action	Count	Share of all pairs
Win	**345**	**70.7%**
Tie	21	4.3%
Loss	122	25.0%

- 488 paired next actions from 100 Qwen BrowseComp trajectories.
- Fixed prompt and summary; anonymized GPT-5.6-Sol comparison.
- Candidate reasoning withheld; counterfactual calls not executed.

Outcome measured: judged quality of a proposed next action.

Paper, Sec. 3.3, PDF p. 6. Criteria were fixed from preceding history.

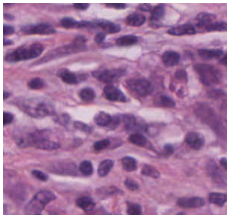
Fewer total compactions accompany fewer responses



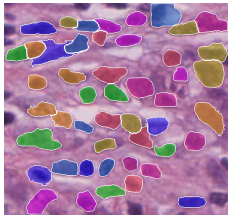
- Paper-trading responses: 5,336 → 3,013; compactions: 56 → 45.
- Per 1,000 responses, compactions rise: 10.5 → 14.9.
- The trace and interval statistics describe different time ranges.

Paper, Fig. 4 and Sec. 3.4, PDF p. 6. Left: Sept. 22 trace. Right: Sept. 20–23 within-session intervals; two residual intervals lie above the plot range but enter the statistics.

The repair case combines memory with file retrieval



Task image



Observed mask overlay

- A refactor deleted helpers; the summary retained the error and backup identity.
- The actor reread both files, restored helpers, and passed all nine tests.
- The trajectory does not identify which recovered details came from memory.

Paper, Fig. 5 and Sec. 3.4, PDF pp. 6–7. Original PNG images, unchanged.

REMORY connects explicit checkpoints and learned states

Existing direction	Role in the paper's positioning
Summaries + retrieval	Retain readable state and access exact records
Continuous compression	Represent longer inputs with learned soft tokens
Distribution matching	Judge compression by the actor's continuation behavior
Recurrent memory	Carry updated state across successive boundaries

Distinctive combination

Summary-conditioned memory, a frozen actor, two-stage training, and recurring compaction boundaries.

Paper, Sec. 4, PDF p. 7. This slide summarizes the paper's related-work positioning.

The useful target is information for future behavior

1. **Representation:** keep a readable summary and train a continuous supplement.
2. **Evidence:** attribution improves, agent scores rise, and fixed-checkpoint next actions are more often preferred.
3. **Trade-off:** final memory is bounded, but encoding and memory-network execution add unpriced computation.

A matched, summary-independent memory baseline would help isolate the value of summary conditioning. (Discussion proposal.)

Synthesis of Secs. 2–5, PDF pp. 3–7. Next-step proposal is reading analysis.

Primary reference and supporting material

Hanchen Xia, Baoyou Chen, Yutang Ge, Naihao Deng, Senqiao Yang, Zilong Dong, Weihao Yuan, and Siyu Zhu.

REMORY: Learning Residual Memory for Context Compaction.
arXiv:2610.11287v1, October 8, 2026.

arxiv.org/abs/2610.11287v1 Versioned paper PDF

- Equations and all quantitative evidence are from this version.
- Source locators use PDF page numbers, matching printed pages.
- Backup slides: gradient path, metric interpretation, and case statistics.

Backup: frozen weights still allow input gradients

Let $M \in \mathbb{R}^{k \times d}$ be the memory vectors, with actor width d.

$$\nabla_\phi \mathcal{L} = \left(\frac{\partial M}{\partial \phi}\right)^\top \sum_j \left(\frac{\partial q_j}{\partial M}\right)^\top \frac{\partial \mathcal{L}}{\partial q_j}.$$

- The teacher distribution p_j supplies a fixed target.
- Student predictions depend on memory through the frozen actor.
- No update to θ is needed to optimize the input interface.

$$D_{\mathrm{KL}}(q_j \| p_j) = \sum_{v \in \mathcal{V}} q_j(v) \log \frac{q_j(v)}{p_j(v)}$$

Teaching derivation from Eqs. (1)–(3), PDF pp. 3–4. d and k introduced for dimensions.

Backup: the three SummHay metrics use different averages

- **Coverage and joint score:** average over all 621 reference insights.
- **Citation F1:** average over covered insights.
- **Shared subset:** attribution still improves on 553 insights covered under both summary conditions.

Coverage and attribution move in opposite directions

Relative to raw context, summary + memory has coverage -4.51, citation F1 $+4.43$, and joint score -0.02 points.

The reported $+4.04$ citation gain over summary uses unrounded values; subtracting displayed values gives $+4.05$.

Paper, Table 1 and Sec. 3.1, PDF pp. 4–5. Differences to raw context calculated from displayed values.

Backup: the trading record mixes several changes

Record	Without memory	With memory
Settled ROI	−11.86%	+5.11%
Logged generation cost	$303.81	$170.30
Responses	5,336	3,013
Completed compactions	56	45
Compactions / 1,000 responses	10.5	14.9
Mean live positions	105,372	101,185

Two concurrent paper-trading accounts; initial bankroll $20,000.
ROI uses resolved markets at the September 23 cutoff. Market choice and timing differ between the two trajectories.

Paper, Sec. 3.4, PDF p. 6. Live position count includes residual memory.