Just-in-Time Memory

Learning to Curate Task-Adaptive Memory for LLM Agents

Paper authors: Yefan Zhou, Yang Li, Zeyu Leo Liu,
Semih Yavuz, Shafiq Joty

Salesforce AI Research

Independent paper-reading presentation

arXiv:2609.27334v1 | 23 September 2026
Reading date: 27 September 2026

Preserve experience; curate it when the next task is known.

The next task determines which details matter

Write-time curation

Finish a task
↓
Commit to a fixed summary
↓
Retrieve it for a future task

Useful details may disappear before their value is known.

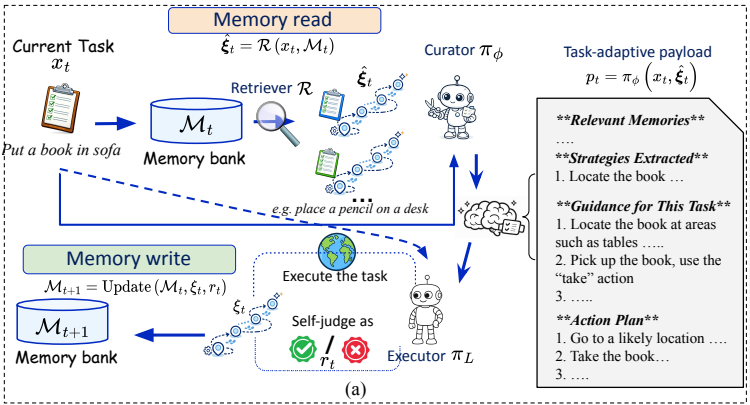
Read-time curation

Store a raw trajectory
↓
Observe the new task
↓
Retrieve and produce tailored guidance

One past experience can support different future lessons.

Paper, Sec. 1, PDF pp. 1–2. Diagram is a teaching reconstruction.

A temporary briefing connects memory to action



Paper, Fig. 1(a), PDF p. 2; Sec. 3, p. 4. Original figure panel, cropped.

- Retrieve raw traces.
- Curate for task x_t.
- Execute with payload p_t.
- Judge and store the new trace.

The payload is not stored.

Raw storage and selective admission are distinct

$$\hat{\xi}_t = \mathcal{R}(x_t, \mathcal{M}_t), \qquad p_t \sim \pi_\phi(\cdot \mid x_t, \hat{\xi}_t)$$

- **Bank $\mathcal{M}_t$:** complete task–observation–action traces.
- **Retriever $\mathcal{R}$:** BM25 over task descriptions; top $k = 3$.
- **Curator π_ϕ:** relevant experience, reusable strategy, current-task guidance.
- **Admission:** the executor model judges whether a completed trace succeeded.

Scope of preservation

Accepted traces remain unabstracted. Rejected traces are discarded.

Paper, Sec. 3, PDF p. 4; Appendix A, p. 18.

The current task directly evaluates its own briefing

For one task, sample $G = 8$ payloads; run the frozen executor on each.

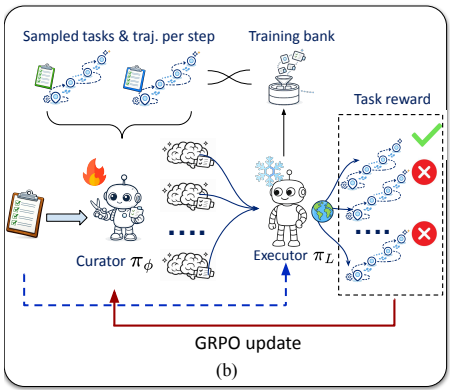
$$\hat{A}_i = r_t^{(i)} - \frac{1}{G}\sum_{j=1}^{G} r_t^{(j)}$$

$$\mathcal{L}_{\mathrm{GRPO}} = -\frac{1}{G}\sum_{i=1}^{G} \hat{A}_i \, \log \pi_\phi\Big(p_t^{(i)} \mid x_t, \hat{\boldsymbol{\xi}}_t\Big)$$

- $r_t^{(i)}$: environment reward; $\hat{A}_i$: centered group advantage.
- No advantage standard-deviation normalization; no value network.
- Credit is immediate *across tasks*; execution still takes multiple actions.

Paper, Sec. 3, PDF pp. 4–5; equations are unnumbered. Table 7 additionally specifies clipping and KL loss.

Only the curator learns from task rewards



- Start from a Qwen3-8B curator, with thinking disabled.
- Train with a frozen Qwen3-8B executor and a **fixed** training bank.
- Use binary success on ALFWorld; continuous score on WebShop.
- 100 updates; batch 32; learning rate 10^{-6}.

No curator RL experiments on τ^2-bench.

Paper, Fig. 1(b), p. 2; Sec. 4.1, p. 7; Table 7, p. 19.

Training uses a fixed bank; testing starts empty

	Curator training	**Default evaluation**
Initial bank	Training-set base traces	Empty bank
Admission	Ground-truth success	LLM judge
Bank updates	Fixed bank	After each test batch
Weights	Curator learns	Frozen
True reward	Train the curator	Score results only

A batched streaming protocol

Tasks within a batch share the same bank state. Task order and batch size can affect performance.

Paper, Sec. 3, PDF p. 5. The training bank does not carry into default testing.

Three benchmarks test different kinds of reuse

Benchmark	Task	Metric ↑	Protocol
ALFWorld	Text-based embodied control	SR	140 tasks; 3 runs
WebShop	Product purchase with constraints	Score, SR	500 tasks; 3 runs
τ^2-bench	Airline, retail, telecom dialogues	Domain and average SR	4 runs; prompted curators only

- Batch sizes: 10 / 10 / 5; main retrieval count: $k = 3$.
- Reported uncertainty: standard deviation across task orderings.
- WebShop score includes partial credit; SR measures successful tasks.

Paper, Sec. 4, p. 5; Tables 1–2, pp. 6–7; Appendix A, p. 18.

Keep curator capacity separate from executor capacity

Comparison methods

- No Memory: executor alone.
- ReasoningBank: written strategies.
- MemP: multilevel procedural items.
- SkillOS: learned skill curation at write time.

Variant names

- `-base`: prompted Qwen3-8B.
- `-gemini/-gpt`: prompted stronger curator.
- JitMem: RL-trained Qwen3-8B curator.

Executors: Qwen3-8B, Gemini-2.5-Pro, or GPT-5.4.

Paper, Secs. 2 and 4, PDF pp. 3, 5. Descriptions of prior work follow this paper.

With Qwen, learned read-time curation wins clearly

Qwen3-8B executor and curator base model; 3 runs.

Method	ALFWorld SR	WebShop score	WebShop SR
No Memory	47.9±1.2	33.3±0.7	9.8±0.5
SkillOS (RL)	61.2±4.6	40.6±0.7	16.5±0.7
JitMem-base	60.5±2.6	32.5±3.2	11.7±0.5
JitMem (RL)	**77.4**±2.9	**61.1**±0.9	**32.8**±1.7

Against the trained write-time baseline

ALFWorld: +16.2 percentage points. WebShop SR: +16.3 points.

Selected rows from paper Table 1, PDF p. 6. Mean ± SD; metrics shown on a 0–100 scale. Baseline provenance: backup slide 23.

An untrained curator helps in some configurations

WebShop SR, matched curator and executor models.

Models	ReasoningBank	SkillOS, prompted	JitMem, prompted
Qwen3-8B	11.4	13.6	11.7
Gemini-2.5-Pro	40.2	41.0	**61.0**
GPT-5.4	33.6	26.9	**47.5**

- Read-time design yields large gains with stronger prompted curators.
- Qwen is a counterexample to universal training-free superiority.
- With Gemini as executor, prompted Gemini beats the RL Qwen curator: 61.0 vs. 50.5.

Paper, Table 1, PDF p. 6. Selected means; see the original table for all standard deviations.

A trained curator transfers across executors

ALFWorld SR; curator base model is Qwen3-8B.

Curator training	Qwen test executor	GPT-5.4 test executor
No RL	60.5±2.6	79.3±3.6
Train with Qwen	77.4±2.9	86.7±0.7
Train with GPT-5.4	—	88.1±0.9

Evidence for modular reuse

Transferring the curator reaches within 1.4 SR points of training it with the target executor.

This experiment keeps the benchmark fixed; it does not establish cross-domain transfer.

Paper, Table 3, PDF p. 7; mean ± SD.

On τ^2-bench, the largest gain is in telecom

GPT-5.4 executor; GPT-5.4 prompted curators; no RL.

Method	Airline	Retail	Telecom	Macro	Micro
No Memory	65.0	79.8	50.2	65.0	65.0
ReasoningBank	65.0	84.6	61.6	70.4	71.7
SkillOS-gpt	68.0	78.5	53.3	66.6	66.4
JitMem-gpt	63.5	84.0	**72.6**	**73.4**	**75.6**

- Against ReasoningBank: Telecom $+11.0$; micro average $+3.9$ points.
- No comparable advantage on Airline or Retail.
- Micro: JitMem 75.6 ± 2.3 vs. ReasoningBank 71.7 ± 3.9.

Paper, Table 2, PDF p. 7; means over 4 task orderings. Macro weights domains equally; micro aggregates tasks.

Executor efficiency improves; total cost is unmeasured

ALFWorld with GPT-5.4 executor, average per task.

Method	Input tokens (K)	Output tokens (K)	Steps
No Memory	9.0	1.40	17.8
ReasoningBank	19.7	1.26	16.2
SkillOS-base	22.4	1.36	16.9
JitMem-base	10.9	1.00	13.2
JitMem	**9.8**	**0.87**	**11.6**

Input tokens fall 50.3–56.3% relative to the two write-time methods.

Executor-side totals exclude curator, judge, retrieval, storage, and training costs.

Paper, Table 4 and accompanying text, PDF p. 7. Lower is better in all three columns.

Task conditioning, filtering, and raw details all help

Untrained Qwen curator and Qwen executor; SR.

Variant	ALFWorld	WebShop
Full JitMem-base	60.5±2.6	11.7±0.5
Remove current task from curator	57.4±2.0	7.1±0.1
Store all traces with outcome labels	59.0±2.2	8.9±1.7
Store write-time distilled items	58.8±1.5	4.9±0.7

- The largest WebShop drop comes from discarding raw details.
- Removing the task from the curator still leaves task-based retrieval.
- One-factor ablations do not establish additive or independent effects.

Paper, Fig. 2(a), p. 8; exact values from Table 9, p. 21. Mean ± SD over 3 runs.

RL gains depend strongly on retrieved experience

Qwen executor; trained Qwen curator unless marked base.

Variant	ALFWorld SR	WebShop SR
Full JitMem	77.4±2.9	32.8±1.7
No retrieved trajectories	62.6±1.2	17.6±1.1
No task conditioning	66.0±4.7	22.4±0.4
Untrained JitMem-base	60.5±2.6	11.7±0.5

Supported conclusion

Retrieval removal costs 14.8 / 15.2 points. Some gains over the untrained model remain; the table supports a narrower claim than the prose.

Paper, Sec. 4.2 and Fig. 2(b), p. 8; Table 9, p. 21. The paper does not fully specify retraining for every RL ablation.

Bank refresh adds modest gains; warm-starting does not

WebShop SR with the trained Qwen curator.

Test executor	Default	Refresh + 50 steps	Warm-start 100 traces
Qwen3-8B	32.8	35.6	32.5
Gemini-2.5-Pro	50.5	50.5	50.9
GPT-5.4	45.4	46.3	44.1

- Refresh changes both the training bank and the training budget.
- No consistent positive effect from preloading the test bank.
- Early-task accuracy is not broken out, so cold-start costs remain unclear.

Paper, Table 5, p. 8; discussion on pp. 8–9. Values are means; complete uncertainty is in the notes.

Qualitative examples reveal procedural guidance

Task: examine a bowl with a desk lamp	
Untrained curator	**RL-trained curator**
Find the bowl; issue a generic examination action.	Take the bowl, move to the desk lamp, and examine the bowl using it.

- Fig. 3: one past trace yields state-change or placement guidance.
- Appendix B: shopping attributes and telecom diagnostic order are reused across instances.
- Examples illustrate behavior; they do not guarantee that every generated instruction is correct.

Paper, Figs. 3–4, pp. 9–10; example payloads, pp. 22–24. Qualitative comparison paraphrased for readability.

Several boundaries matter for interpretation

Authors' limitations

- Simple BM25 retrieval.
- An extra curator call per task.
- Benchmark-specific payload format.
- Curation only once per task.

Reading observations

- No full-system cost comparison.
- Judge accuracy is not quantified.
- Some baseline values are inherited.
- 3–4 orderings are not independent training-seed evidence.

Paper, Sec. 5, p. 10; Table 4, p. 7; baseline reproduction, Table 6, p. 18. Right column is reader analysis.

Defer abstraction when future uses are uncertain

1. **Architecture:** store accepted raw experience, then curate a temporary task-specific briefing.
2. **Learning:** same-task reward trains a modular curator without waiting for later tasks.
3. **Evidence:** strong benchmark gains and executor transfer, with important model and protocol dependence.

A practical follow-up experiment

Compare raw retrieval, prompted curation, and RL curation under the same total token and latency budget.

Synthesis of paper Secs. 3–5, pp. 4–10. Follow-up experiment is a reader proposal, not a reported result.

Source and reading scope

Yefan Zhou, Yang Li, Zeyu Leo Liu, Semih Yavuz, Shafiq Joty.

Just-in-Time Memory: Learning to Curate Task-Adaptive Memory for LLM Agents.
Salesforce AI Research, 2026.

`https://arxiv.org/abs/2609.27334v1`

- Read the author TeX, all main sections, and Appendices A/B.
- Checked key equations, tables, figures, and locators against the 25-page PDF.
- Figures reuse author assets; tables select and reformat reported values.
- This presentation analyzes the paper; it does not reproduce its experiments.

Companion Chinese notes contain full explanations, uncertainty, and source locators. PDF page numbers match printed pages in v1.

Backup: the implementation includes clipping and KL

Setting	Reported value
Curator / training executor	Qwen3-8B, non-thinking
Updates / batch / group	100 / 32 prompts / 8 samples
Learning rate / warmup	10^{-6} / 5 steps
Clipping / loss aggregation	$(0.2, 0.2)$ / token-mean
KL loss	Enabled, coefficient 10^{-3}
KL in reward	Disabled
Prompt / response cap	32768 / 8192 (ALFWorld), 4096 (WebShop)
Hardware / reported run time	8 H200s; about 21 h / 27 h

Paper, Appendix A and Table 7, PDF pp. 18–19. Test-time Qwen executor: thinking on ALFWorld, non-thinking on WebShop.

Backup: no-memory baseline provenance needs care

Qwen3-8B, reported vs. reproduced means.

Source	ALFWorld SR	WebShop score	WebShop SR
SkillOS reported / main Table 1	47.9	33.3	9.8
Reproduced, non-thinking	34.5	36.4	8.6
Reproduced, thinking	42.1	29.0	4.4

- Executor prompts and thinking settings were calibrated during reproduction.
- A stronger reported SR is conservative for that gap, but not proof of identical configurations.
- Table 6's "all at or below" description has a score exception: 36.4 exceeds 33.3.

Paper, Tables 1 and 6, PDF pp. 6 and 18. Values are transcribed from the paper; observations are reader analysis.

Backup: sensitivity claims have a limited scope

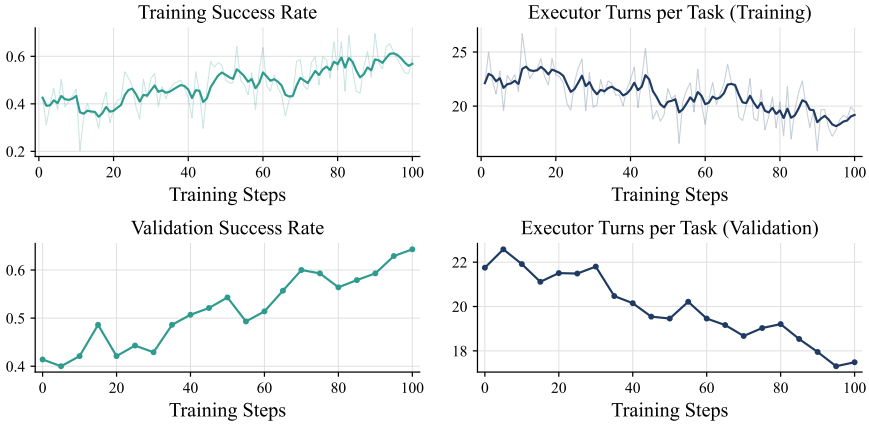
WebShop SR, number of retrieved traces.

Executor	$k = 3$	$k = 5$
Qwen3-8B	32.8±1.7	32.8±0.3
Gemini-2.5-Pro	50.5±0.8	48.6±0.7
GPT-5.4	45.4±0.0	45.3±0.8

- Evidence covers two small k values; larger banks are not tested here.
- Warm-starting GPT yields 44.1±0.4 vs. 45.4±0.0 by default.
- That 1.3-point gap is not literally within the displayed SDs; a paired significance test is not reported.

Paper, Table 8, p. 20; Table 5, p. 8 and prose on p. 9. Mean ± SD; no confidence intervals are implied.

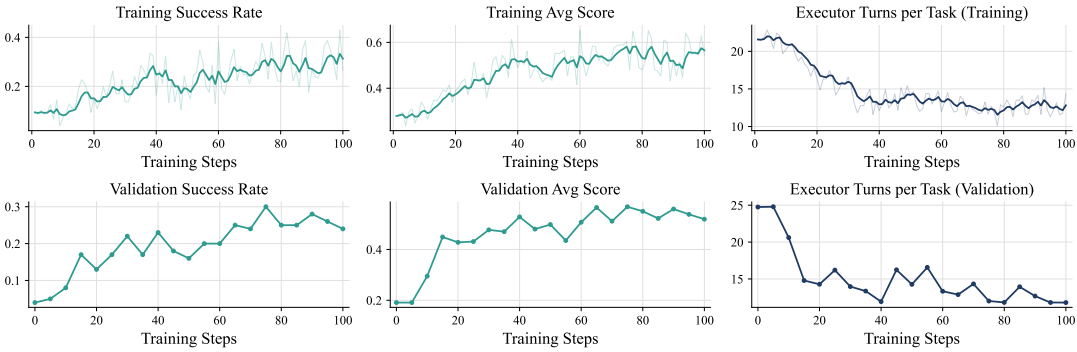
Backup: ALFWorld improves over 100 training steps



Top: training. Bottom: validation. Success rises and executor turns fall overall.

Original paper Fig. 5, PDF p. 24. Validation uses a subsampled test stream; its role in model selection is not fully specified.

Backup: WebShop improves with visible fluctuations



Columns: success rate, average score, executor turns.

Rows: training and validation. Trends are not strictly monotonic.

Original paper Fig. 6, PDF p. 25. These plots do not quantify variability across independent RL training runs.