A Gain Is Not Proof of Learning: SkillRC, a Reality Check for Distilled Agent Memory

Yiwen Lu
University of Notre Dame
ylu37@nd.edu

Research report, November 2026. Code and data: https://github.com/EvenEureka/SkillRC

Abstract

We introduce SkillRC, a reality check for distilled agent memory that compares each memory with matched texts and repeats the comparison over pools built by the same pipeline. A matched control keeps a memory's items, order and token count and replaces its content, so a gain over no memory splits into a reading effect (control minus no memory) and a content value (memory minus control). With six language models on ALFWorld and WebShop, the measured value of distilled rules is modest: across four local models and many pools, the estimated advantage of a memory over unrelated text is 1.2 to 5.0 points. Over three pools, the value depends on the comparison: distilled rules lead generic rules from the same writer by 3.5 points but advice for another environment by only 1.1 [−2.7, 4.9]. Replicate pools vary on the scale of the value itself, with a standard deviation of 1.9 points, and one WebShop pool lowers scores by 11.6 to 15.8 points against its control. Adding the control also changes which gains an evaluation detects: 8 of 16 net gains on single pools come with no detected content gain. A gain on one memory is therefore weak evidence of learning, and the unit worth evaluating is the pipeline that builds memories.

1 Introduction

Agents reuse past experience by distilling successful runs into a short list of rules placed in the prompt (Zhao et al., 2024; Shinn et al., 2023; Wang et al., 2025). The evidence that such a memory works is almost always a gain over the same agent without memory, and these gains are uneven: they vary across tasks and models, are sometimes negative (Li et al., 2026; Han et al., 2026), and shrink when skills have to be retrieved under realistic conditions (Liu et al., 2026). In contrast, we show that a gain over no memory is weak evidence that the agent learned from experience. Against matched text, the measured value of distilled rules is modest, depends on what it is compared with, and varies between pools from the same pipeline on the scale of the value itself.

Concretely, we introduce SKILLRC, a reality check (RC) for distilled agent memory (Figure 1). Adding a memory changes two things at once: the model has more text to read, and the text says something about the task. Language models are sensitive to irrelevant context and to the length and format of their input (Shi et al., 2023; Levy et al., 2024; Sclar et al., 2024), and across six pools the gardening control described below lowered Llama-3.1-8B's success by 4.4 points relative to no memory. SkillRC therefore compares a memory with a *matched control*: the same items with the same tags and word counts, in the same order and with the same token count under each tested model's tokenizer, but with each rule replaced by a sentence about vegetable gardening. The control minus no memory is the *reading effect*. The memory minus the control is the *content value*. SkillRC then asks what this value is made of, by comparing the memory with rules written without experience and with advice for another environment, and how far it holds, by repeating the comparison over replicate pools built by the same pipeline.

We apply SkillRC on ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022) with six models, starting from a 33-rule list selected because it helped GPT-4o-mini: it raised that model's success on ALFWorld by about 9 points and changed the success of four other models by less than 3. Across many pools, the estimated advantage of a memory over unrelated text is 1.2 to 5.0 points for four local readers. Over three pools and two Qwen3 readers, distilled rules lead generic rules from the same writer by 3.5 points and advice for another environment by 1.1. Replicate pools vary about as much as the value itself for three Qwen3 readers, one WebShop pool lowers scores by 11.6 to 15.8 points for two readers, and scrambling the

rules lowers success by 5.3 points for GPT-4o-mini and 8.0 for Qwen3-8B (Sections 5.1–5.4).

SkillRC is a matched-control design for agent memory whose match to what the harness inserts is verified on every task for each reader, extended with comparison texts and replicate pools (Section 3). It also changes what an evaluation detects: adding the control changes the detection verdict in 11 of 54 evaluations, and the gap between two estimates of one reader is 54 to 69% smaller with eight pools than with one (Section 5.5). For claims that a gain comes from experience, we recommend evaluating the pipeline that builds memories over several pools and more than one control.

2 Related Work

From storing experience to evaluating what it adds. Agent memory methods turn past runs into reusable instructions, and their representations have grown from single reflections into curated libraries. They fall into three families: executable skills (Wang et al., 2024), natural-language lessons such as reflections, distilled insights and evolving playbooks (Shinn et al., 2023; Zhao et al., 2024; Fu et al., 2024; Zhang et al., 2026), and reusable workflows (Wang et al., 2025). Later systems curate skills around a fixed executor (Yang et al., 2026; Ouyang et al., 2026; Gao et al., 2026) or optimize the whole harness in which a memory sits (Lee et al., 2026; Huang et al., 2026), and most of this work supports its methods with a gain over the same agent without memory. Evaluations that look across tasks and models find such gains unstable. SkillsBench reports negative deltas on some tasks and notes that an injected skill also adds context, using self-generated skills as a partial control (Li et al., 2026), and SWE-Skills-Bench finds limited gains with a token overhead (Han et al., 2026). Self-EvoSkill compares executions with and without a skill to revise the skill (Gao et al., 2026). Along this line, evaluation has moved from one gain to gains across tasks and models. SkillRC asks how a gain changes when the comparison text and the sampled pool change.

From prompt controls to repeated memory evaluations. Prompting research has long tested what remains of a prompt's effect when its content is replaced. Content-free inputs calibrate few-shot predictions (Zhao et al., 2021), some irrelevant instructions work about as well as good ones in the cited experiments (Webson and Pavlick, 2022), and demonstrations with random labels keep much of their benefit (Min et al., 2022). On ALFWorld itself, ReAct performance is driven by the similarity between demonstrations and queries more than by the content of the reasoning traces (Verma et al., 2024). These controls were built for a single prompt. An agent's memory is reread at every step of an episode by models whose tokenizers differ, and it comes from a pipeline that builds lists from different trajectory samples. SkillRC brings replacement controls to agent memory: it builds the control for one specific memory, audits the match to what the harness inserts on every task for each reader, and repeats the comparison over pools from one pipeline.

3 Method

SkillRC runs the same tasks under three conditions that differ only in one block of the prompt: a memory, a matched control or nothing (Figure 1). A memory is a pool of short rules, each tagged with a task type and distilled from successful trajectories. We call the frozen model together with its tokenizer, chat template and serving the *reader*.

3.1 Reading effect and content value

A comparison with and without a memory changes both the amount of text and what it says. A third condition compares the memory with text of the same shape and different content. Let $\bar{Y}_{\mathsf{none}}$, $\bar{Y}_{\mathsf{mem}}$ and $\bar{Y}_{\mathsf{ctrl}}$ be a reader's mean success without memory, with the memory and with the control. The net effect splits as

$$\underbrace{\bar{Y}_{\mathsf{mem}} - \bar{Y}_{\mathsf{none}}}_{T,\ \text{net effect}} = \underbrace{\bar{Y}_{\mathsf{ctrl}} - \bar{Y}_{\mathsf{none}}}_{I,\ \text{reading effect}} + \underbrace{\bar{Y}_{\mathsf{mem}} - \bar{Y}_{\mathsf{ctrl}}}_{R,\ \text{content value}}. \tag{1}$$

T is what a with-and-without comparison reports. A negative I means that a block of unrelated text lowers success, and a positive R means that the rules score above that text in the same place. We use the names reading effect and content value as labels for these two measured differences, and Section 5.2 shows how far they move when the control changes.

3.2 Matched control

Length, format and position all change what a model does, so the control matches the memory in each of them. For each tested reader, we call two conditions *payload-matched* when they insert corresponding items in the same order and with

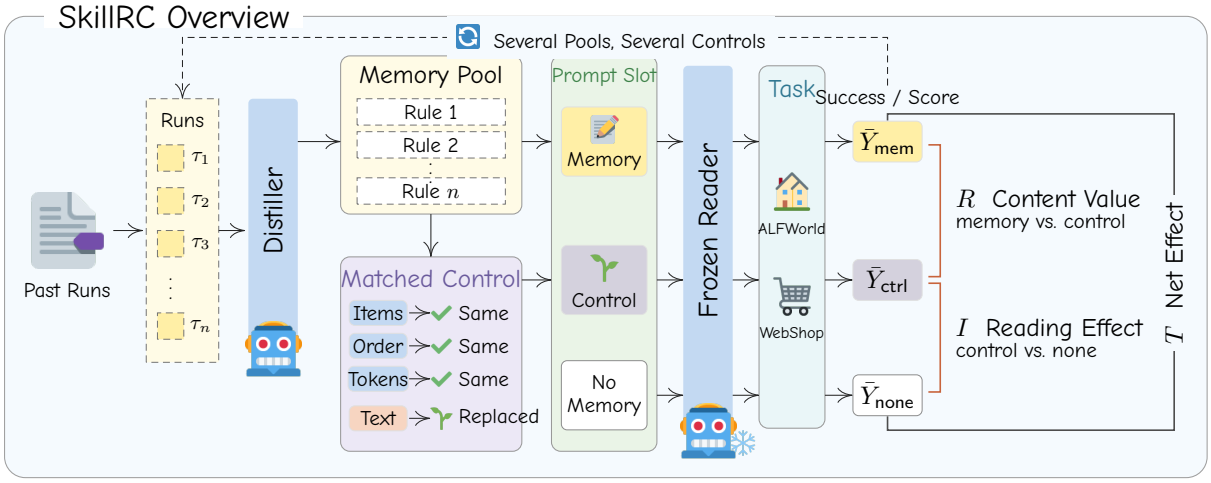


Figure 1: **SkillRC evaluates a memory against matched text and across replicate pools.** One frozen reader runs the same tasks three times, and only the text in one prompt slot changes: a distilled memory, a matched control with the memory's items, order and token count but replaced text, or nothing. The gaps between the three success rates give the content value R and the reading effect I, which add up to the net effect T that a with-and-without comparison reports. The audit is repeated over several pools and several controls.

equal token counts on every task. Given a memory, we first keep each item's tag and word count and replace its body with an imperative sentence about vegetable gardening, drawn from a bank screened against a vocabulary exclusion list; the first control retains some incidental overlap, including the word "fork". For example, "Always clean the item using the designated cleaning area before placing it in the final location" becomes "Always pull weeds while they are small so they never set seed across the whole garden", which also has 16 words. We then rank the control with the memory's BM25 scores (Robertson and Zaragoza, 2009), so that each item lands in the same position. Finally, a greedy pass, deterministic first and then with seeded orders, replaces body words with the nonce token x until the block has the memory's token count.

We verify the match on what the harness actually inserts. The harness logs the token count in every episode, and an offline audit replays retrieval on every task to confirm the items and their order.

3.3 Comparison texts and replicate pools

The matched control shows what the rules add over unrelated text. Two further texts of the same shape ask where that addition comes from. Generic rules are written by the same model without seeing a trajectory, so they show what the writer says without experience. Advice for another environment keeps a realistic rule-list format and carries strategies written for a different task, so it tests a task-specific advantage.

A pool is also one draw from a pipeline. SkillRC therefore reports estimates pooled over replicate pools, which keep the source agent, the distiller and the number of trajectories per task type and change only the seed that selects the trajectories. With k pools and one control an audit needs $2k + 1$ runs per reader, against 2 for one with-and-without comparison, so eight pools take 17 runs, and each further control adds k.

All estimates are differences of paired task means. Intervals are 95% percentile intervals from 20,000 bootstrap draws of tasks within task type (product category on WebShop), also resampling subsamples within source for the crossover and individual pools or reader-pool cells for the same-window pooled analyses.

4 Experimental Setup

Environments. On ALFWorld we use the 134 unseen games with a ReAct scaffold (Yao et al., 2023) that lists the admissible actions, two demonstrations per prompt and a limit of 30 steps; each run covers every game with three demonstration pairs, 402 episodes. WebShop replaces rooms and objects with search and product pages, which tests whether the decomposition carries over to another action space. We use the 1,000-product environment with a seeded server and 500 test sessions of at most 15 steps, scored in $[0, 1]$. In our WebShop harness the observations do not name the current page.

Readers. The six readers span a range of ability, from 26% to 90% success on ALFWorld without memory. GPT-4o-mini and GPT-4.1-mini are called through the OpenAI API. Llama-3.1-8B (Grattafiori et al., 2024) and Qwen3-8B, Qwen3-14B and Qwen3-32B (Yang et al., 2025), with thinking off, are served locally with vLLM (Kwon et al., 2023). All decode at temperature zero.

Memories. Our main evaluation is a crossover at balanced coverage. Each of three source agents (GPT-4o-mini, Qwen3-8B, Qwen3-32B) supplies the same number of successful trajectories per task type, three replicate subsamples per source, and each subsample is distilled once by Qwen3-32B and once by Qwen3-8B. ALFWorld trajectories come from the 140 seen games and WebShop's from 576 training sessions, and evaluation uses separate test sets. Subsamples from one source share some trajectories, because a source has a limited number of successes. This gives 18 pools of 31 to 36 rules, each with its own gardening control, and the three Qwen3 readers read all of them, for 54 evaluations. Llama-3.1-8B reads six of these pools. GPT-4o-mini reads the 33-rule pool from the introduction, distilled from its own successes. On WebShop one pool of 30 rules, distilled from GPT-4o-mini's successes, is compared with a control about choral singing, because WebShop sells garden products.

Baselines. The baseline evaluation is the net effect T on one pool. On three pools in one time window, Qwen3-8B and Qwen3-14B also read two further texts of the same shape: generic rules that Qwen3-32B, the writer of these pools, wrote without seeing a trajectory, and advice for another environment, rephrased from our WebShop rule pools (for example, "Scan listing title on the page for the requested size before clicking at all") and screened against the control builder's exclusion vocabulary, which exempts stopwords. A separate study compares gardening and astronomy controls on two pools with Qwen3-8B and Qwen3-14B.

Protocols. These evaluations come from six studies whose predictions, estimands and analyses were written down before the runs, except for one addendum written after two hosted readers had been analysed. Together the studies cover 24 ALFWorld pools. We mark everything chosen after seeing data as exploratory. Repeating a run under identical settings changes 17–20% of GPT-4o-mini episodes and 4–14% of local-reader episodes.

5 Results

5.1 Against unrelated text, a memory's value is modest

Across four local readers, the estimated advantage of a memory over unrelated text is 1.2 to 5.0 points (Table 1). Over many pools the content value is $+5.0$ points for Llama-3.1-8B and $+3.9$ for Qwen3-8B, both with intervals above zero, and $+1.3$ $[-0.8,\ 3.4]$ and $+1.2$ $[-0.8,\ 3.3]$ for Qwen3-14B and Qwen3-32B. The largest value, $+9.1$, is GPT-4o-mini's on a single pool built from its own successes, the pool we chose because it had helped that model.

Net gains and gains over matched text give different pictures of the same memory. For Llama-3.1-8B over six pools, gardening text scores 4.4 points below no memory and the memory scores 5.0 above gardening, for a net of $+0.6$ $[-2.9,\ 4.0]$ with no detectable spread between pools, so a with-and-without comparison detects no gain for this reader. In every one of these six pools gardening scores below both other conditions. An exploratory comparison on three pools in one window finds the same pattern for Qwen3-8B, with shifts of -2.8 and $+5.4$, while for Qwen3-14B neither shift excludes zero. Similar nets also come with different content values: Qwen3-14B's net of $+3.3$ is close to Qwen3-8B's $+4.5$, yet their content values are $+1.3$ and $+3.9$, and only Qwen3-8B's interval excludes zero.

5.2 The measured value depends on the comparison

The reported estimate moves with the comparison condition and with the pools averaged (Table 2). Without a control, Llama-3.1-8B's $+5.0$ reads as $+0.6$ and Qwen3-14B's $+1.3$ reads as $+3.3$. With one pool in place of many, a single estimate for Qwen3-8B falls anywhere from -1.5 to $+10.4$.

The memory leads generic rules by 3.5 points and advice for another environment by 1.1. Over the same three pools and both readers, the memory leads gardening by $+3.1$ $[-0.8,\ 7.7]$, generic rules from the same writer by $+3.5$ $[0.2,\ 7.2]$, as we predicted, and advice for another environment by $+1.1$ $[-2.7,\ 4.9]$. We had registered a lead over that advice as well, and the test did not detect it. The generic rules themselves score -0.4 $[-3.0,\ 2.3]$

Table 1: **Estimated memory advantages over matched unrelated text are 1.2 to 5.0 points across four local ALFWorld readers, while the tested WebShop pool lowers scores by 11.6 to 15.8 points for two readers.** The top row is what a with-and-without comparison reports, and the shaded rows are the decomposition, in points of success ($\times 100$; WebShop: score), with 95% intervals. Bold marks an interval that excludes zero. Llama-3.1-8B: six pools. Qwen3 readers: 18 pools. GPT-4o-mini on ALFWorld: one pool, mean of two runs. WebShop: one pool.

	ALFWorld					WebShop	
	Llama-3.1-8B	Qwen3-8B	Qwen3-14B	Qwen3-32B	GPT-4o-mini	GPT-4o-mini	Qwen3-8B
Net effect T (with vs. without)	+0.6 [−2.9, 4.0]	**+4.5** [0.9, 8.0]	**+3.3** [0.03, 6.8]	+0.2 [−2.1, 2.7]	**+8.8** [5.0, 12.8]	**−14.0** [−17.4, −10.6]	**−13.2** [−17.1, −9.3]
Reading effect I (ours)	**−4.4** [−8.2, −0.7]	+0.6 [−2.9, 4.2]	+2.0 [−1.4, 5.5]	−1.0 [−3.8, 1.8]	−0.2 [−3.7, 3.1]	+1.8 [−1.2, 4.8]	−1.6 [−5.6, 2.4]
Content value R (ours)	**+5.0** [2.2, 8.0]	**+3.9** [1.2, 6.5]	+1.3 [−0.8, 3.4]	+1.2 [−0.8, 3.3]	**+9.1** [5.0, 13.3]	**−15.8** [−19.1, −12.5]	**−11.6** [−15.3, −7.9]

Table 2: **The reported value changes with the comparison condition and with the pools included.** Each row is one version of the evaluation, in points of success. Many pools: six for Llama-3.1-8B, 18 for Qwen3, against gardening text. Generic rules come from the same writer as the pools, and other-domain advice is written for WebShop. The highlighted row is the matched gardening control over many pools.

Version	Llama-3.1-8B	Qwen3-8B	Qwen3-14B
Gardening control	+5.0	+3.9	+1.3
No control	+0.6	+4.5	+3.3
One pool (range)	[+3.7, +6.5]	[−1.5, +10.4]	[−2.0, +6.2]
Three pools, one time window: memory against			
Gardening text	–	+5.4	+0.8
Generic rules	–	+5.1	+1.9
Other-domain advice	–	+1.7	+0.4

against gardening, and one reader-pool cell supplies nearly half of the memory's lead over them.

The choice of unrelated text matters too. In a separate study on two pools with Qwen3-8B and Qwen3-14B, replacing gardening with a second unrelated topic, amateur astronomy, moves the control's score by -2.5 points $[-4.4, -0.7]$. The measured reading effect therefore differs between these two control texts.

Coherent rules outperform their scrambled words for two readers. Replacing the rules with their own words in scrambled order lowers success by 5.3 points for GPT-4o-mini and 8.0 for Qwen3-8B, so these readers respond to more than the rules' vocabulary. The estimated source and distiller contrasts are -0.2 $[-2.4, 1.9]$ (Qwen3-8B against GPT-4o-mini) and $+0.5$ $[-1.2, 2.2]$ points.

5.3 Replicate pools vary on the scale of the value

Pools built the same way differ about as much as the value itself. For a fixed source, distiller and reader, replicate pools differ by more than task sampling explains ($Q = 72.2$ on 36 degrees of freedom, $p = 0.0003$), with a between-pool standard deviation of 1.9 points, about the size of the content values themselves. For Qwen3-8B, three pools distilled by Qwen3-32B from GPT-4o-mini's successes give 0.0, $+6.0$ and $+10.4$. For the net effect and the content value alike, the detection verdict on one pool matches the verdict on a replicate pool in only 34 of 54 pairs.

5.4 One WebShop memory lowers scores

The one WebShop pool we tested lowers scores against both no memory and its matched control (Table 1). It takes GPT-4o-mini's score from 29.4 to 15.4 while the choral control leaves it at 31.2, a content value of -15.8 $[-19.1, -12.5]$, and for Qwen3-8B the content value is -11.6 $[-15.3, -7.9]$. We had predicted a positive value for GPT-4o-mini. The loss comes with more searching (exploratory count): the pool raises GPT-4o-mini's searches per session from 2.44 to 5.69, against 2.35 with the control, and several of its rules tell the agent to search again.

5.5 Adding the control changes what an evaluation detects

Half of the net gains that a with-and-without comparison detects come with no detected content gain (Table 3). Of the 16 evaluations with a net gain at the nominal 95% level, 8 detect no content gain against the matched control. Another 3, all for Qwen3-32B, detect a content gain without a detected net gain, for 11 changed verdicts in 54 evaluations. After Holm correction across the 54 evaluations, one net gain and three content gains remain detected.

Averaging over pools shrinks the disagree-

Table 3: **Adding the matched control changes 11 of 54 detection verdicts, and averaging eight pools shrinks the gap between two estimates.** Baseline against SkillRC. Upper rows: gains detected on single pools by the net effect (baseline) and by the content value against the matched control, for three Qwen3 readers on ALFWorld; detection means a 95% task interval above zero. Lower rows: mean absolute gap between two content-value estimates of one reader, with one pool each (baseline) and eight pools each.

	Baseline	SKILLRC (ours)
Adding the matched control: 54 one-pool evaluations		
Gains detected	16	11
Gains detected by both designs	8	8
Gains only this design detects	8	3
Adding pools: gap between two estimates, in points ↓		
Qwen3-8B (one pool vs. eight)	3.8	**1.8**
Qwen3-14B (one pool vs. eight)	2.4	**0.8**
Qwen3-32B (one pool vs. eight)	2.3	**0.7**

ment that one pool leaves. Each eight-pool side averages both distillations of four subsamples, chosen from the nine available across three sources. Two estimates of one reader differ by 2.3 to 3.8 points with one pool each and by 0.7 to 1.8 with eight, a gap 54 to 69% smaller. Comparisons across sources, whose two sides share no trajectories, point the same way: over the three pairs of sources, the mean gap per reader is 0.2 to 2.2 points for six pools each, against 2.3 to 3.7 for single pools.

6 Conclusion

A gain over no memory is the usual evidence that an agent learned from experience, and SkillRC shows that it is weak evidence. Against matched unrelated text, the estimated advantage of distilled rules is 1.2 to 5.0 points for four local readers, and against advice written for another environment it is 1.1. Pools from one pipeline vary about as much as the value, one WebShop pool lowers scores by 11.6 to 15.8 points, and adding the control changes 11 of 54 detection verdicts. Scrambling the rules lowers success by 5.3 points for GPT-4o-mini and 8.0 for Qwen3-8B, and on three pools the rules beat generic rules from the same writer by 3.5 points. For claims that a gain comes from experience, we recommend reporting a memory against at least two matched controls and over several pools, and treating the pipeline that builds memories as the unit of evaluation. The next tests are long procedural skills under a binding retrieval budget and advice from a third domain.

Limitations.

- *What the contrasts mean.* Reading effect and content value are relative to the chosen control. Reading them as the effect of reading and the effect of content assumes that reading costs the same whatever the text is about, which the astronomy control does not support, and the benefit of exact token matching and placeholder tokens has not been tested against simpler controls.
- *Coverage.* The memories are rule lists of 520 to 790 tokens placed whole in the prompt, so retrieval never chose among items, and the readers are four small open models and two small hosted models. The generic-rule and other-environment comparisons ran on three pools and two readers, the astronomy control on two pools, the generic rules all derive from one base list, the hosted readers read only a control with placeholder tokens, and the WebShop result is one pool in a harness whose observations do not name the current page.
- *Run variation.* Intervals exclude variation across runs and time windows, and the 1.9-point spread between pools includes run variation that we did not estimate separately. These intervals do not establish whether differences of about 2 points between individual runs exceed run variation.
- *Shared data.* Replicate pools share trajectories, on average 38% of the smaller side in the eight-pool comparisons, so those gaps are partly optimistic, and the cross-source comparison that avoids overlap compares different sources. The crossover shares one no-memory run per reader, the six Llama-3.1-8B pools derive from three subsamples, and Llama-3.1-8B's two contrasts reuse the same gardening runs.
- *No ground truth.* We have no ground truth for a memory's value, so SkillRC changes detection verdicts and narrows the gap between estimates without showing that decisions made with it are more accurate. A nondetection, such as the lead of $+1.1$ over other-domain advice, does not establish that a value is zero.

References

Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. 2024. AutoGuide: Automated generation and selection of context-aware guidelines for large language model agents. In *Advances in Neural*

Information Processing Systems, volume 37, pages 119919–119948.

Haowen Gao, Haoran Chen, Can Wang, Shasha Guo, Liang Pang, Zhaoyang Liu, Huawei Shen, and Xueqi Cheng. 2026. SelfEvoSkill: Paired execution audits for skill revision without external outcome supervision. *Preprint*, arXiv:2606.14239.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, et al. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tingxu Han, Yi Zhang, Wei Song, Chunrong Fang, Zhenyu Chen, Youcheng Sun, and Lijie Hu. 2026. SWE-Skills-Bench: Do agent skills actually help in real-world software engineering? *Preprint*, arXiv:2603.15401.

Yue Huang, Wenjie Wang, Han Bao, Yuchen Ma, Xiaonan Luo, Yi Nian, Haomin Zhuang, Zheyuan Liu, Yue Zhao, and Xiangliang Zhang. 2026. MemoHarness: Agent harnesses that learn from experience. *Preprint*, arXiv:2607.14159.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. 2026. Meta-Harness: End-to-end optimization of model harnesses. In *Conference on Language Modeling (COLM)*.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Xiangyi Li, Yimin Liu, Wenbo Chen, Bingran You, Zonglin Di, Yifeng He, Shenghan Zheng, Kyoung Whan Choe, Jiankai Sun, Shuyi Wang, Chujun Tao, Binxu Li, Xuandong Zhao, Hejia Geng, Xiaojun Wu, Junwei Zhou, Xiaokun Chen, Hanwen Xing, Yubo Li, et al. 2026. SkillsBench: Benchmarking how well agent skills work across diverse tasks. *Preprint*, arXiv:2602.12670.

Yujian Liu, Jiabao Ji, Li An, Tommi Jaakkola, Yang Zhang, and Shiyu Chang. 2026. How well do agentic skills work in the wild: Benchmarking LLM skill usage in realistic settings. In *Conference on Language Modeling (COLM)*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siru Ouyang, Jun Yan, Yanfei Chen, Rujun Han, Zifeng Wang, Bhavana Dalvi Mishra, Rui Meng, Chun-Liang Li, Yizhu Jiao, Kaiwen Zha, Maohao Shen, Vishy Tirumalashetty, George Lee, Jiawei Han, Tomas Pfister, and Chen-Yu Lee. 2026. SkillOS: Learning skill curation for self-evolving agents. *Preprint*, arXiv:2605.06614.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR)*.

Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2024. On the brittle foundations of ReAct prompting for agentic large language models. *Preprint*, arXiv:2405.13966.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. 2025. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 63897–63911. PMLR.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, et al. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, Kai Qiu, Yuqing Yang, Dongdong Chen, Xue Yang, and Chong Luo. 2026. SkillOpt: Executive strategy for self-evolving agent skills. *Preprint*, arXiv:2605.23904.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, et al. 2026. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations (ICLR)*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. ExpeL: LLM agents are experiential learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19632–19642.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Implementation Details

Models. The GPT-4o-mini alias resolved to the snapshot of 2024-07-18, and GPT-4.1-mini was called as the snapshot of 2025-04-14. Local models were served with vLLM in bf16 with a context of 16,384 tokens, with `enable_thinking=false` for Qwen3. Evaluation calls used at most 256 completion tokens, and distillation calls at most 512.

Estimators. For the between-pool standard deviation, we average $((R_i - R_j)^2 - v_{ij})/2$ over replicate pairs within each source-distiller-reader cell and then across cells, truncate the pooled variance at zero, and report its square root; v_{ij} is the task-bootstrap variance with a small-sample correction.

Pools. Demonstration pairs are selected by seeds 0, 1 and 5. Trajectories are balanced at six per task type except one for heat, because GPT-4o-mini solved only one heat game, 31 in all. Pools are distilled with one call per task type at temperature 0, from a prompt that asks for 3 to 6 concise general rules, each one imperative line with no object numbers or room layouts. Generic rules use the same instruction with a fixed description of the environment in place of trajectories, and each is then rewritten to the matched item's word count. The focal control needed the most placeholder tokens, 30 of 472 body words; the crossover controls needed none.

Artifacts. The repository on the first page holds the harness (`skillrc/`), configurations (`configs/`), pools and result summaries (`data/`) and analysis scripts (`scripts/analysis/`). The crossover is under `configs/w1_2026_10_13` and `data/summaries/w1a_2026-10-13`. The same-window configurations and summaries are not yet public; that study uses replicate 1 of each source, the Qwen3-32B distillations for the three-pool comparisons and both distillations for Llama-3.1-8B's six pools, with the same demonstration seeds. Figure icons are from Twemoji (CC BY 4.0).