Composing Transferable Skills and Verbal Reflection Improves Agent Performance

Peter Lubell-Doughtie *peter@ona.io*
Ona

Abstract

Building dynamic context to orchestrate actors in large language model systems is an active area of research essential to producing high-quality results on long-horizon tasks. However, combining the different and effective research methods being pursued and tested is not straightforward. We propose rSSO, a composition of skill-library learning (SSO) and verbal reflection (Reflexion). rSSO maintains two distinct memory systems, one with skills distilled from successful trajectories and the other with reflections drawn from failed trajectories. Evaluated in the text environment ScienceWorld, rSSO achieves performance neither method achieves alone. rSSO outperforms SSO by 8.95 and Reflexion by 6.28 points.

1 Introduction

Large language model (LLM) agents in interactive environments learn from experience without weight updates by accumulating in-context memory across attempts. Two prominent families of methods occupy this design space. Skill libraries (Nottingham et al., 2024; Wang et al., 2023; Zhao et al., 2024) extract reusable action sequences from successful trajectories, where each skill is abstracted into an instruction list and a target state. The skills are retrieved per-step at run time via embedding-based similarity to the current observation. Verbal reflection (Shinn et al., 2023; Majumder et al., 2024; Madaan et al., 2023) compiles summaries of what went wrong from failed trajectories. These summaries are then prepended to the actor's prompt on subsequent trials. The two modalities draw from complementary signals, success versus failure, and inject information at complementary points in the prompt, per-step retrieval versus once-per-trial preamble.

Here we consider whether skill libraries and verbal reflection compose. Their composability would provide us with an opportunity for performance gains by combining SSO with Reflexion, and potential future work integrating further methods. If they do not compose, the two modalities are either redundant or adversarial in some characterizable way. We resolve this question empirically on ScienceWorld (Wang et al., 2022) using Claude Sonnet 4.6 (Anthropic, 2026) as the base model. Our contributions are as follows:

- **Composition of SSO with Reflexion (rSSO).** SSO and Reflexion use distinct memory aggregation methods, per-step skill retrieval and once-per-trial reflection preamble, but the existing literature does not measure them together. Across four different learning algorithms (ReAct, Reflexion alone, SSO alone, and their composition rSSO) under the SSO adaptation setting, rSSO outperforms all others, scoring 8.95 points above SSO and 6.28 points above Reflexion (see Table 1).
- **Claude Sonnet results on the ScienceWorld benchmark.** To our knowledge no prior published numbers on this benchmark use a Claude model. SSO alone on Sonnet 4.6 reaches 79.46, 4.24 points below Nottingham et al. (2024)'s GPT-4 number of 83.7, which we treat as faithful-within-tolerance reproduction (see §4). rSSO also exceeds the published GPT-4 number at 88.41, but the base-model difference between Sonnet 4.6 (2026) and GPT-4 (2023) is the dominant contributor to that cross-model gap. The within-scaffold increase reported above is the claim we rely on.
- **Evidence characterizing where composition helps.** The increased performance concentrates on task types where SSO alone plateaus mid-run. In these cases, reflection contributes a verbal

account of why the plateau persists that per-step skill retrieval alone does not surface (Fig. 2, Table 2).

2 Background

2.1 ScienceWorld and the SSO Adaptation Setting

ScienceWorld (Wang et al., 2022) is a text-based science simulation with 30 task types spanning physics, chemistry, biology, and genetics. Each task has multiple variants that use different target substances, locations, or thresholds. The actor must navigate rooms, manipulate objects, and ultimately register a result via a `focus on` action. Scoring is defined in the $[-100, 100]$ range with positive partial credit for sub-goals, -100 for committing to a wrong answer, and 100 for correct completion.

SSO (Nottingham et al., 2024) evaluates two settings, *adaptation* and *transfer*. We follow its adaptation setting on an 18-task subset of ScienceWorld's 30 task types. For each task, the actor attempts each of 10 held-out test variants with 5 trials per variant. Learning is allowed between trials within a variant but the actor memory is cleared between variants. Under this setting with GPT-4 SSO scores 83.7, ReAct (Yao et al., 2023) 29.6, and Reflexion (Shinn et al., 2023) 39.4.

2.2 Skill Set Optimization

SSO maintains a per-variant skill library. After each trial, the trajectory is split at positive-reward boundaries with sub-trajectories of length $\in [3, 6]$ steps as candidates for skill construction. Cross-trajectory matching (requires ≥ 2 trajectories with the same window) identifies recurring action patterns, which are then scored on coverage, reward, state similarity, and action similarity. Beam search selects a non-overlapping set of skills, with each selected skill verbalized by a 3-stage prompt (summary, instructions, and target state). At run time, the actor receives the top-3 skills per step by cosine similarity between the skill's stored initial state and the actor's current state.

2.3 Reflexion

Reflexion maintains a per-task FIFO buffer of natural-language reflections (default size 3). After each failed trial (env score < 100), the entire trajectory is summarized by an LLM into a single reflection string emphasizing what went wrong and what to try differently. On the next trial, all buffered reflections are prepended to the actor's prompt as a `Plans from past attempts` block before the task description. No reflections are generated on success.

2.4 Why these have not been composed

The SSO results compare against Reflexion as a baseline and report SSO outperforming Reflexion by a wide margin on adapt (83.7 vs. 39.4). The presentation positions Reflexion as a weaker alternative, although the two methods' memory layouts (per-step skill retrieval into the action-template region of the prompt and once-per-trial reflection prepended to the system message) suggest they are composable. To the best of our knowledge, no published study has tested them together. The closest adjacent work, ExpeL (Zhao et al., 2024), extracts insights as rules across multiple tasks, rather than per-step skills, and has not been evaluated on ScienceWorld.

3 Method

We compose SSO and Reflexion by maintaining two distinct memories that survive across the within-variant trials. The two memories are written by separate LLM calls, on separate triggers, and read into non-overlapping regions of the actor prompt. Figure 1 shows how the rSSO actor builds its dynamic prompt by composing two distinct memory aggregation methods.

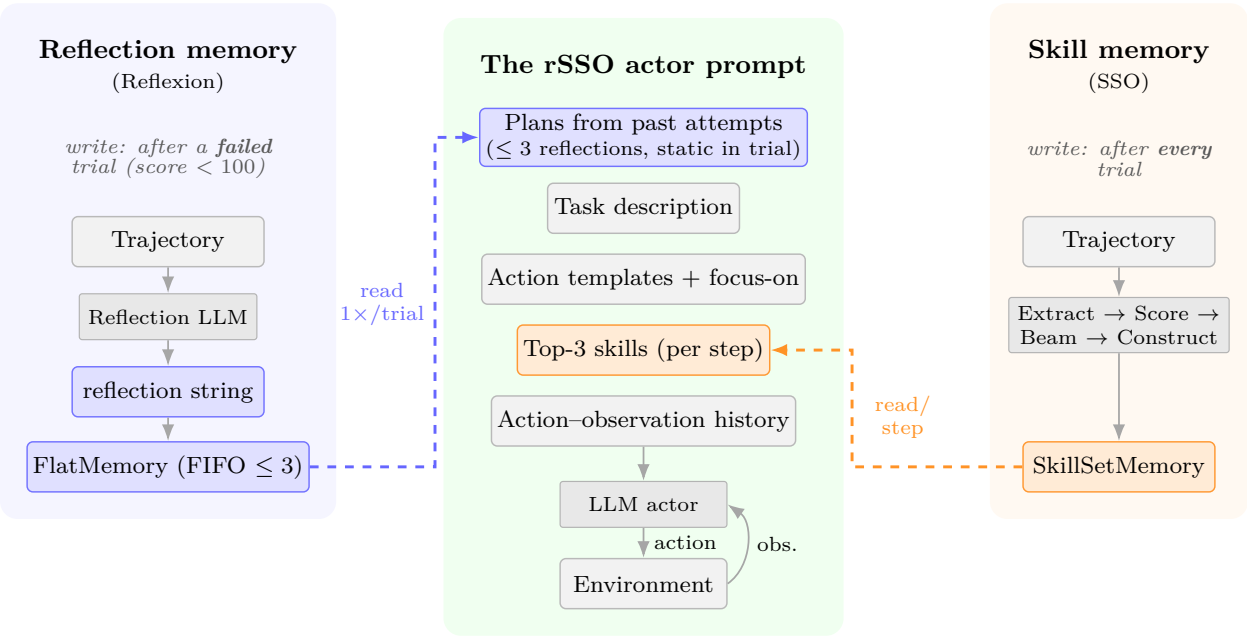

Figure 1: **rSSO composes two distinct memory aggregation methods.** *Center:* the actor prompt at run time. The reflection preamble (blue) and the per-step skill region (orange) occupy non-overlapping positions. All gray regions are inherited unchanged from pure SSO. *Left:* Reflexion's reflection memory is written by a single LLM call after a failed trial and read once as the preamble (loaded at trial start, static through the trial). *Right:* SSO's skill memory is written after every trial by the cross-trajectory pipeline and read per step into the skill region. The two memory stores draw on disjoint signals, successes for skills and failures for reflections, and never read or write the same prompt region.

1. A **SkillSetMemory** (SSO's data structure) holding trajectories, candidate skill windows, and the current beam-searched skill set. This data structure is updated after every trial using sub-trajectory extraction, cross-trajectory matching, scoring, beam search, and 3-stage LLM skill construction.
2. A **FlatMemory** (Reflexion's data structure) — a FIFO queue of up to N=3 verbal reflections. Updated after every failed trial (env score < 100) by appending a single LLM-generated reflection summarizing the failure.

At each step, the prompt is assembled as in Algorithm 1.

Algorithm 1 Per-step prompt assembly for the rSSO actor; $\|$ denotes concatenation.

Require: FlatMemory R, SkillSetMemory K, current state s

$P \leftarrow \langle\,\rangle$ *// prompt, assembled top to bottom*
$P \leftarrow P \,\|\, R$ *// plans from past attempts*
$P \leftarrow P \,\|$ task description
$P \leftarrow P \,\|$ action templates and focus-on warning
for $k \in \textsc{TopSkills}(K, s, 3)$ **do** *// top-3 by cosine similarity to s*
    $P \leftarrow P \,\|$ instruction list and target state of k *// instructions for reaching subgoals*
end for
$P \leftarrow P \,\|$ action–observation history
return P

Reflections are static within a trial, i.e. they are loaded once at trial start, while skills are retrieved fresh on every step, matching SSO's per-step-retrieval behavior. rSSO includes SSO's 3-stage verbalization of

each newly constructed skill, run after every trial, and adds one reflection call per failed trial. The actor's structured-output contract, its reasoning, current subgoal, and action, is unchanged from SSO.

4 Experimental Setup

Protocol. We follow the SSO adaptation setting (Nottingham et al., 2024) consisting of 18 task types, 10 test-split variants per task, 5 trials per variant, with the actor's memory cleared between variants. Following SSO, max steps per trial is 1.5 × the gold-action-sequence length, rounded down and dynamic per variant. We select the test variants following SSO's selection process. Of the $18 \times 10 = 180$ nominal variants, 164 are available in the test split (some tasks have < 10 test variants).

Evaluation metric. ScienceWorld returns a positive-reward accumulator score in $[0, 100]$ and an environment score in $[-100, 100]$ that reaches 100 only on a correct solution. We report three metrics over these, following SSO. The *Mean* is computed from the accumulator score, taken per variant as the best across the 5 trials and averaged over the 164 variants ($\text{mean}_v \max_t$, where v indexes the variant and t the trial). *Solve* is the fraction of variants reaching environment score ≥ 100 on any trial, and *Acc100* the fraction reaching accumulator score ≥ 100 on the best trial.

Model. For our base model we use Claude Sonnet 4.6 (Anthropic, 2026) for both the actor and reflection/skill-construction.

Embeddings. We use all-MiniLM-L6-v2 (Reimers & Gurevych, 2019) for skill-retrieval similarity. This deviates from SSO's text-embedding-ada-002, which we discuss further in §6.

Implementation faithfulness. We verified our SSO implementation against the SSO reference implementation (Nottingham et al., 2024), matching its 3-stage skill generation, cross-trajectory matching, beam search, and LLM-based deduplication. We did not implement SSO's optional learned skill-scoring feedback loop as it does not materially affect the reported number per the SSO paper.

Prompt scaffold inheritance. Each method in our comparison uses the prompt scaffold described by that method. The SSO and rSSO rows use the SSO action-template prompt with structured current-subgoal output and state preprocessing, as released by Nottingham et al. (2024). The ReAct and Reflexion rows use the ReAct-style prompt from Shinn et al. (2023). This matches the comparison setting the SSO paper itself uses to position SSO against Reflexion. A consequence is that trial 1 of each method, in which no learned memory has yet been acquired, is not identical across rows. Under our setting with empty memories, SSO reaches 75.30 in trial 1 while Reflexion reaches 58.37. We interpret this gap as evidence that SSO's contribution includes the prompt-scaffold design as a static prior, in addition to the skill-retrieval mechanism. The composition score increase we report (rSSO − SSO, +8.95) is unaffected as both rows share the SSO scaffold and differ only in the presence of reflection memory.

5 Results

rSSO outperforms both SSO and Reflexion. rSSO reaches a mean of 88.41, the best of the four methods on all reported metrics (Table 1), and its score distribution dominates the others at every score threshold (Fig. 3). rSSO exceeds SSO by 8.95 points (79.46 to 88.41) and Reflexion by 6.28 (82.13 to 88.41). Looking across task types, rSSO matches or exceeds SSO on 16 of the 18 task types, a per-task-type improvement significant at $p \approx 0.009$ in the two-sided Wilcoxon signed-rank test (details in Appendix A).

The composition outperformance is monotone across trials. The per-trial cumulative best-of-t means in Fig. 2 show that rSSO is ahead of SSO alone at every $t \geq 2$, with 84.49 vs. 76.63 at t=2, 85.80 vs. 77.57 at t=3, 87.55 vs. 78.09 at t=4, and 88.41 vs. 79.46 at t=5. The gap opens at t=2, when reflection memory first carries information, and stays open. SSO alone makes small per-trial gains of +1.33, +0.94, +0.52, and +1.37, while rSSO has a large trial 1 to 2 increase of +9.42 and smaller subsequent gains.

Method	Scaffold†	Mean	Solve	Acc100
ReAct	flat	58.37	62/164	56/164
Reflexion	flat	<u>82.13</u>	<u>114/164</u>	<u>102/164</u>
SSO	SSO	79.46	113/164	96/164
rSSO	SSO	**88.41**	**134/164**	**112/164**
ReAct	flat	29.6	—	—
Reflexion	flat	39.4	—	—
SSO	SSO	83.7	—	—

Table 1: ScienceWorld best-of-5 results on the SSO adaptation setting (N=164 test variants). Top panel uses Sonnet 4.6 as the LLM, and rSSO exceeds SSO alone by +8.95 points using the same scaffold and code. Within the top panel **best** is in bold and <u>second-best</u> underlined. Bottom panel shows existing results using GPT-4 as the LLM.

†Each method inherits the prompt scaffold published with it, see §4. Within-scaffold deltas isolate the memory configuration, while cross-scaffold or cross-panel deltas have confounders. The ReAct row is trial 1 of the Reflexion run, which has empty reflection memory, making it identical to a ReAct baseline.

Where the performance improvement concentrates. Per-task inspection shows the score increase over SSO is not uniform across the 18 task types (full breakdown in Appendix A). Tasks where SSO alone plateaus at a sub-100 score on every trial, where skill retrieval finds a partial procedure but the actor never closes the last step, are the ones reflection rescues most often. Reflection adds a verbalized account of the last failure that gives the actor a missing piece of context the per-step skill prompt does not surface (Appendix B). Tasks SSO alone already solves at t=1 ($\approx$ 60% of variants) cannot gain from reflection because no failure transcript is ever produced.

Reflection's performance improvement transfers across scaffolds. The trial 1 to 2 performance improvement is large in both scaffold contexts, +10.87 for Reflexion-alone, +9.42 for rSSO. The two scaffolds reach very different absolute levels at t=1 (Table 1), but the marginal effect of adding verbal reflection is comparable. This is consistent with skill libraries and verbal reflection drawing from non-overlapping signals, positive-reward sub-trajectories in the former and failure summaries in the latter, then using these at different positions in the dynamically constructed prompt.

Reflection memory more than closes the scaffold gap. In the flat scaffold, adding reflection memory raises the score from 58.37 to 82.13, an increase of 23.76 points; this is substantially larger than the within-SSO-scaffold composition increase of 8.95 points. Reflexion-alone (82.13) modestly exceeds SSO-alone (79.46) by +2.67 on the same setting. We caution against over-reading this cross-scaffold comparison. Scaffold and memory both differ between these two rows, and the within-scaffold increases above are the controlled comparisons. Still, the pattern is consistent with a stronger base model making more of the verbal-reflection channel and less of SSO's static prompt-scaffold prior. The Reflexion/GPT-4 result reported by Nottingham et al. (2024) on this setting is 39.4, suggesting the model-version increase from GPT-4 to Sonnet 4.6 is over 42 points on Reflexion alone, enough to largely close the reported Reflexion-vs-SSO gap of 44.3 points.

6 Limitations

Our study focused on composition and does not vary the base model. All numbers use Claude Sonnet 4.6, and the cross-model row in Table 1 is included as landscape context, not a controlled comparison. We also report only ScienceWorld. Whether rSSO transfers to other interactive-agent benchmarks (ALFWorld, WebShop, TAU-Bench) is an open question. Because the two memory aggregation methods do not interfere, we expect the composition to transfer, which we leave to future work.

Additionally, the empty-memory trial 1 gap between SSO and Reflexion (75.30 vs. 58.37, §4) bounds the prompt-scaffold contribution from above but does not separate it from the skill-retrieval contribution. The

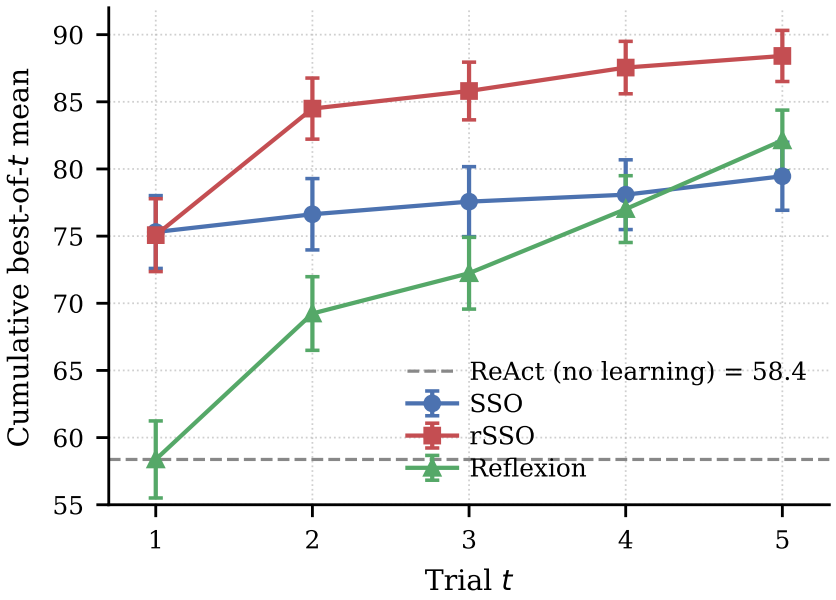

Figure 2: Per-trial cumulative best-of-t mean across 164 task variants on the SSO adaptation setting with Sonnet 4.6, computed for each variant as the best score over trials 1 through t (of the 5 trials) and averaged over variants. Error bars are SEM across variants. rSSO (red) leads at every $t \geq 2$. Reflexion-alone (green) crosses SSO-alone (blue) at t=4 and finishes +2.67 points above it at t=5 (82.13 vs. 79.46); the ReAct baseline (58.37, dashed) is reported as t=1 of the Reflexion run since reflection memory is empty at t=1.

full decomposition, running ReAct and Reflexion under the SSO scaffold with empty skill libraries, remains open.

Finally, our reproduction makes three implementation choices that favor comparability across SSO, Reflexion, and rSSO over matching the published SSO/GPT-4 numbers. We use MiniLM-L6-v2 for skill-retrieval embedding rather than text-embedding-ada-002, we do not implement SSO's optional learned skill-scoring feedback loop, and we use Reflexion's default buffer of 3 reflections per task without ablating it. Each of these choices is held fixed on both sides of the comparison it could affect, so it can shift the absolute scores but not the relative differences we report.

7 Conclusion

We asked whether two in-context learning methods, skill libraries (Nottingham et al., 2024) and verbal reflection (Shinn et al., 2023), compose and found that in the ScienceWorld environment they do. Their composition, rSSO, outperforms both constituents, by 8.95 points over SSO and 6.28 over Reflexion, reaching a level neither method attains on its own. rSSO achieves this without any new components, keeping two memories, one of skills distilled from successful trajectories and one of reflections drawn from failures. The performance improvement concentrates on the task types where SSO plateaus, and a verbal account of the last failure supplies context that per-step skill retrieval does not surface. These results point to a broader hypothesis for future work, that distinct memory aggregation methods drawing on complementary signals can be composed to improve agent performance.

References

Anthropic. Introducing sonnet 4.6, February 2026. URL `https://www.anthropic.com/news/claude-sonnet-4-6`. Model release.

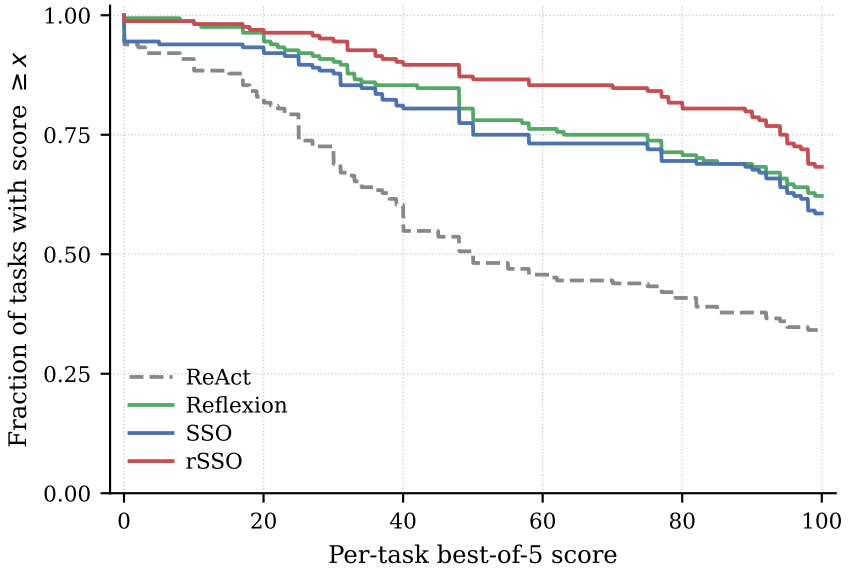

Figure 3: Survival function of per-task best-of-5 scores across the four Sonnet 4.6 conditions, where $S(x) = P(\text{score} \geq x)$ is the fraction of the 164 test variants that reached at least score x. A higher curve at each x means more tasks reached at least that score. The score on the x-axis is the positive-reward accumulator (same signal as the *Mean* column in Table 1), so the rightmost value of each curve (x=100) equals the Acc100 column of Table 1, not the Solve column. rSSO (red) dominates all other methods at every score. Reflexion (green) and SSO (blue) are intertwined through the middle of the range, with Reflexion edging out SSO from $x \approx 70$ upward. ReAct (dashed) sits well below all three.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL `https://arxiv.org/abs/2303.17651`.

Bodhisattwa Prasad Majumder, Bhavana Dalvi Mishra, Peter Jansen, Oyvind Tafjord, Niket Tandon, Li Zhang, Chris Callison-Burch, and Peter Clark. CLIN: A continually learning language agent for rapid task adaptation and generalization. In *Conference on Language Modeling (COLM)*, 2024. URL `https://arxiv.org/abs/2310.10134`.

Kolby Nottingham, Bodhisattwa Prasad Majumder, Bhavana Dalvi Mishra, Sameer Singh, Peter Clark, and Roy Fox. Skill set optimization: Reinforcing language model behavior via transferable skills. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL `https://arxiv.org/abs/2402.03244`.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. URL `https://arxiv.org/abs/1908.10084`.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL `https://arxiv.org/abs/2303.11366`.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. In *Transactions on Machine Learning Research*, 2023. URL `https://arxiv.org/abs/2305.16291`.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022. URL `https://arxiv.org/abs/2203.07540`.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL `https://arxiv.org/abs/2210.03629`.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. URL `https://arxiv.org/abs/2308.10144`.

A Per-task-type breakdown

Table 2 reports the per-task-type mean best-of-5 ScienceWorld score (the same accumulator signal used in the *Mean* column of Table 1) for each of the four Sonnet 4.6 conditions, plus the within-scaffold deltas. Numbers are means over N test variants per task type; the bottom row recovers the overall means from Table 1 (N=164).

Task type	N	ReAct	Reflexion	SSO	rSSO	Δ_{SSO}	Δ_{flat}
boil	9	41.3	77.8	64.6	89.8	+25.2	+36.4
chemistry-mix	8	71.2	81.6	82.0	82.0	0.0	+10.4
chemistry-mix-paint-secondary-color	9	30.0	67.8	71.1	86.7	+15.6	+37.8
find-living-thing	10	52.5	88.3	52.5	90.0	+37.5	+35.8
find-plant	10	100.0	100.0	91.7	91.7	0.0	0.0
freeze	9	45.8	59.6	48.9	63.2	+14.3	+13.8
grow-fruit	10	20.8	34.7	63.9	77.0	+13.1	+13.9
grow-plant	10	45.7	71.1	77.9	78.2	+ 0.3	+25.4
identify-life-stages-1	5	61.6	100.0	95.4	82.0	−13.4	+38.4
identify-life-stages-2	4	46.5	46.5	68.5	71.5	+ 3.0	0.0
inclined-plane-determine-angle	10	44.0	99.0	99.0	93.0	−6.0	+55.0
inclined-plane-friction-named-surfaces	10	87.0	100.0	100.0	100.0	0.0	+13.0
lifespan-longest-lived	10	62.5	95.0	87.5	100.0	+12.5	+32.5
lifespan-shortest-lived	10	60.0	95.0	77.5	95.0	+17.5	+35.0
measure-melting-point-known-substance	10	61.7	98.5	80.3	98.5	+18.2	+36.8
mendelian-genetics-known-plant	10	100.0	100.0	100.0	100.0	0.0	0.0
mendelian-genetics-unknown-plant	10	24.5	50.7	68.7	76.6	+ 7.9	+26.2
use-thermometer	10	86.8	96.1	97.3	99.1	+ 1.8	+ 9.3
Overall	164	58.37	82.13	79.46	88.41	+8.95	+23.76

Table 2: Per-task-type mean best-of-5 score on the SSO adaptation setting with Sonnet 4.6. Δ_{SSO} = rSSO − SSO (within-SSO-scaffold composition improvement, by task type) and Δ_{flat} = Reflexion − ReAct (within-flat-scaffold composition improvement). Two task types, `identify-life-stages-1` and `inclined-plane-determine-angle`, show a negative within-SSO-scaffold delta, the only cases where the rSSO composition underperforms SSO alone on this setting. Both show large positive within-flat-scaffold deltas, so reflection memory does help under the ReAct scaffold. The SSO scaffold appears to already extract most of the available signal on those specific tasks. `find-plant`, `mendelian-genetics-known-plant`, and `inclined-plane-friction-named-surfaces` are saturated (≥ 95) across all four methods.

Across the 18 task types, Δ_{SSO} is non-negative on 16 and strictly positive on 12 (four types tie exactly). A Wilcoxon signed-rank test over the 18 per-task-type deltas gives W=12, $p \approx 0.009$ (two-sided). An exact sign test on the 14 non-tied types (12 favoring rSSO) gives $p \approx 0.013$. The largest within-SSO-scaffold performance improvements are on `find-living-thing` (+37.5), `boil` (+25.2), `measure-melting-point-known-substance` (+18.2), and `lifespan-shortest-lived` (+17.5). Two task

types are exceptions. `identify-life-stages-1` and `inclined-plane-determine-angle` show negative Δ_{SSO} (-13.4 and -6.0), but both also show large positive within-flat-scaffold performance improvements ($+38.4$ and $+55.0$). SSO-alone substantially outperforms Reflexion-alone on `grow-fruit` (63.9 vs. 34.7), `mendelian-genetics-unknown-plant` (68.7 vs. 50.7), and `identify-life-stages-2` (68.5 vs. 46.5). On each of these Δ_{SSO} remains non-negative.

B Reflection-driven recovery on a plateau task

To illustrate how reflection memory supplies context that per-step skill retrieval does not, we trace `boil::22`, a variant on which SSO alone plateaus and rSSO recovers; its best-of-t scores across the five trials are $0, 0, 0, 75, 100$. The early failures stem from a recurring error, confusing a tin container (a tin cup) with elemental tin as a substance and searching the kitchen. By trials 4 and 5 the actor's reasoning explicitly cites the accumulated reflections and corrects course.

> *Trial 4.* "Past attempts warn against confusing tin containers with tin substance, and suggest searching chemistry labs or workshops."
>
> *Trial 5.* "Past attempts warn to find tin as a pure substance (not a tin cup), likely in a workshop or chemistry lab, and to use a high-temperature source like a blast furnace."

The corrective content, identify pure tin, search the workshop rather than the kitchen, and use a blast furnace, is a verbal account of why the earlier attempts failed. While the task is still failing, SSO has no successful `boil` sub-trajectory to retrieve, so this context reaches the actor only through the reflection channel.