SERIORA
RESEARCH

cs.AI / cs.LG / cs.SE

MOST SELF-IMPROVING LLM-AGENT PAPERS FAIL TWO OR MORE EVALUATION TRAPS: A METHODS AUDIT OF FORTY-ONE PAPERS

SERIORA RESEARCH PREPRINT

Lionel Arce
Seriora Research
larce@seriora.ai

September 22, 2026

ABSTRACT

Self-improving LLM agents update prompts, memories, tools, or scaffolds from experience and often report large task gains. Published eval protocols leave standard validity threats unchecked, and prior work has not measured multi-trap failure rates over the SI-agent paper population. We freeze 41 peer-visible 2024–2026 SI-agent papers, code each on five traps with two machine-assisted passes from arXiv abstracts and HTML, and human-spot-check 20% of papers. Primary `ge2` is $39/41 = 95.1\%$, which exceeds the pre-registered 50% threshold; ignore-length locked `ge2` is $38/41 = 92.7\%$. This is a methods audit of reported protocols, not a new agent experiment and not a claim that residual SI gain is zero.

Keywords self-improving agents · evaluation threats · LLM agents · methods audit · reproducibility

1 Introduction

Self-improving LLM agents update prompts, memories, tools, or scaffolds from their own experience. Published papers often report large task gains. Those gains are hard to interpret when eval protocols leave standard threats unchecked.

Budget-matched studies already show that reflection-style scaffolds can look strong until tokens are equalized [Wang et al., 2024, Mirzaei, 2026]. Agent-facing work documents fragile variance, order effects, and harness integrity failures [Ye et al., 2026, Bjarnason et al., 2026, Wang et al., 2026c]. Benchmark audits find similar integrity problems in agent eval suites [Zhu et al., 2025]. What remains missing is a population-level multi-trap statistic about the SI-agent literature itself.

This paper supplies that statistic. We audit 41 frozen peer-visible 2024–2026 SI-agent papers with a five-trap checklist. A paper fails the multi-trap bar when at least two applicable traps fail (`ge2`). Under a pre-registered $> 50\%$ rule, $39/41 = 95.1\%$ of primary papers are `ge2`. That share exceeds the pre-registered 50% threshold. Sensitivity that ignores the length trap still leaves locked `ge2` at $38/41 = 92.7\%$ (pass 1 was $36/41 \approx 87.8\%$).

The unit of analysis is the published eval protocol, not a new agent run. We differentiate from equal-token reflection audits by covering persistent SI families and by counting multi-trap prevalence rather than re-ranking one scaffold family.

Contributions.

- A locked population statistic. Primary `ge2` is $39/41 = 95.1\%$ under the pre-registered $> 50\%$ rule, with ignore-length locked at $38/41 = 92.7\%$.
- Per-trap and per-family fail rates on the same frozen sheet, including a locked coding matrix in the appendix.
- A frozen five-trap rubric and a full paper$\times$trap matrix so readers can re-threshold.

2 Related Work

Prior work settles important pieces of the threat model. We did not find a prior literature-majority multi-trap coding of peer SI-agent papers.

2.1 Budget-matched reflection and scaffolds

Token Economies shows that complex reasoning strategies often lose to cost-matched chain-of-thought with self-consistency once query and token budgets align [Wang et al., 2024]. Sample More, Reflect Less sharpens the same point with equal generated-token comparisons. Self-Refine and Reflexion do not beat repeated sampling across a large grid of small and mid-size models on math [Mirzaei, 2026]. These papers validate the budget trap for episode-local reflection. They do not audit persistent memory, harness evolution, prompt compilers, or research agents as a paper population.

Test-time compute scaling further motivates matched baselines [Snell et al., 2025]. Our audit treats budget mismatch as one trap among five, not as the whole claim.

2.2 Variance, order, and randomness

Fragility re-evaluations show that self-improving agents can amplify run variance and hide curriculum effects in default task order [Ye et al., 2026]. Work on randomness in agentic evals argues that small single-run gains are often noise [Bjarnason et al., 2026]. These results support our variance trap. They measure systems under re-run, not the share of published SI papers that already report multi-seed protocols.

2.3 Adjacent majority findings on other units

ABC audits agentic benchmarks and finds widespread validity and reporting issues [Zhu et al., 2025]. BenchJack proposes a long checklist for agent benchmark integrity [Wang et al., 2026a]. Harness tampering audits find nonzero tampering rates across several self-improving systems' trajectories [Wang et al., 2026c]. Anatomy of Agentic Memory taxonomizes agentic-memory evaluation limits and system-cost oversights [Jiang et al., 2026]. That survey is adjacent as a taxonomy of memory-eval failure modes, not a five-trap coding of SI papers as a population. Rethinking harness-evolution evaluation compares automatic harness search to matched test-time baselines and held-out tasks [Wang et al., 2026d]. These are majority-style or protocol findings on benchmarks, runs, memory systems, or harness search. Our unit is the peer-visible SI paper's reported eval protocol under a fixed five-trap checklist.

2.4 Persistent SI systems

Exemplar SI systems include experiential memory [Liu et al., 2025]. Prompt and context compilers appear in DSPy, ACE, GEPA, and TextGrad [Khattab et al., 2024, Zhang et al., 2025b, Agrawal et al., 2025, Yuksekgonul et al., 2025]. Research agents include The AI Scientist line [Lu et al., 2024, Yamada et al., 2025]. Open-ended self-modifying agents include the Darwin Godel Machine [Zhang et al., 2025a]. Preregistration proposals argue that agent experiments still lack locked protocols [Vaccaro, 2026]. We cite these as the literature we audit or as methodological context. We do not re-litigate each system's headline gain here.

2.5 Residual novelty

Equal-token reflection is already contested. Our contribution is different. We measure how often peer SI-agent papers fail at least two of five written traps, with frozen inclusion and dual machine-assisted coding. We did not find a prior literature-majority multi-trap coding of peer SI-agent papers. Anatomy [Jiang et al., 2026] and ABC [Zhu et al., 2025] remain adjacent under different units.

3 Method

This paper is a methods audit of published eval protocols. It is not a new agent experiment and not a re-run bake-off.

3.1 One claim

Under a fixed five-trap checklist, a majority of 2024–2026 peer-visible self-improving LLM-agent papers fail at least two traps. The primary population is papers with at least four applicable traps. The claim exceeds the decision threshold if the primary `ge2` share exceeds 50%. It is falsified if that share is at most 20%.

3.2 Inclusion

We freeze inclusion before coding. A paper is `yes` if it is dated on or after 2024-01-01, claims self-improvement / self-evolution / experiential learning or a persistent memory, prompt, or scaffold update for LLM agents, and reports a quantitative task metric. We exclude pure surveys and weight-only continual learning without an agent loop. Integrity-audit papers that meet this rule stay in the SI corpus. Phantom Guardrails and Auditing Harness Tampering are included as stress tests of the checklist. Adjacent equal-token, budget-control, and scoop-risk *methods* papers that do not claim a new SI system stay in Related Work only (Token Economies, Sample More Reflect Less, ABC, BenchJack). Harness Tampering [Wang et al., 2026c] is dual-cited. It is a corpus member under inclusion and an adjacent finding on runs under a different unit of analysis. The frozen set has 41 `yes` papers. Each paper is assigned one family among memory, prompt-opt, self-mod, reflection-scaffold, ai-scientist, and other.

3.3 Five traps

Coding follows a written rubric. Prefer what the paper states over charitable inference.

held_out. Pass if the SI loop builds the artifact on train or experience data and the primary metric is on held-out tasks. Fail if optimize-and-report uses the same instances with no stated separation.

gate. Pass if persistent updates do not require ground-truth correctness, oracle env reward, or model-judge admission as the write gate. Fail if only GT-correct trajectories are stored, gold answers stop reflection, correctness$\times$efficiency weighted objectives gate admits, or an LLM judge is the admit path during SI. Final offline metrics may still use gold labels.

length_ctrl. `NA` if there is no persistent text artifact. Pass if a length-matched or irrelevant-text control is reported. Fail if persistent memory, playbook, cheatsheet, or prompt text is the claimed mechanism without such a control.

budget. Pass if the primary comparison matches tokens, queries, env steps, or dollars against a non-SI baseline, or if the paper explicitly reports numeric asymmetric cost. Fail if SI is compared only to a cheaper baseline with no matched or asymmetric report.

variance. Pass if at least five seeds or run-level uncertainty over at least five independent runs is reported for the primary metric. Fail otherwise.

3.4 Aggregates

For each paper, `applicable` counts non-`NA` traps. `fails` counts fails among those. `primary` is 1 when `applicable`$\geq$4. `ge2` is 1 when `fails`$\geq$2. All 41 included papers are primary under this rule.

3.5 Coding protocol

Two independent machine-assisted coding passes score all 41 papers from local arXiv abstracts and HTML full text. Pass 1 and pass 2 do not treat each other as ground truth. Cell agreement is $168/205 = 82\%$. Disagreements are adjudicated with written evidence notes. A human spot-check re-reads eight papers (40 trap cells) against the adjudicated sheet and finds 35/40 AGREE. Lionel accepts five flips and locks the sheet. Thirty-three of 41 papers were not human-reread cell by cell. The locked sheet is the sole source of truth for Results. We do not change locked trap cells after the fact.

Table 1: Decision rule and locked primary outcome.

Item	Value
Inclusion (frozen `yes`)	41 papers
Primary filter	applicable ≥ 4
Primary N	41 / 41
Exceeds threshold if	`ge2` share $> 50\%$
Falsified if	`ge2` share $\leq 20\%$
Locked `ge2`	$39/41 = 95.1\%$
Decision	exceeds pre-registered 50% threshold

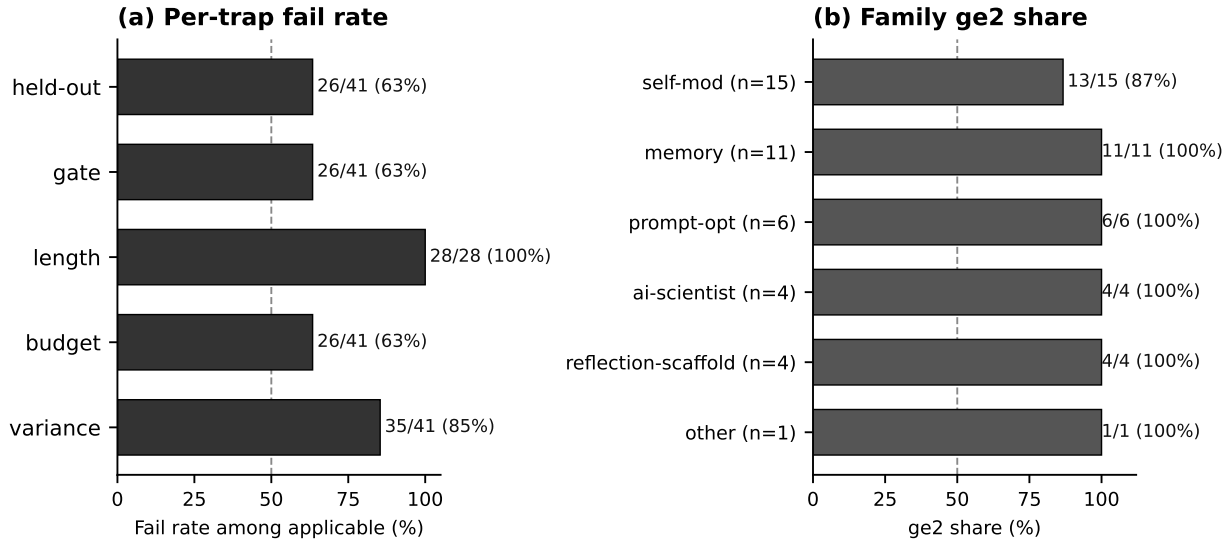

Figure 1: Locked summary ($N = 41$). (a) Fail rate among applicable cells for each trap (dashed line at 50%). (b) `ge2` share by mechanism family. Cell-level coding is in Appendix A.

4 Results

We report locked coding after two independent machine-assisted passes, a 40-cell human spot-check, and five human-accepted flips. The decision rule is fixed in advance. The primary claim exceeds the threshold if the share of primary papers with at least two trap fails (`ge2`) exceeds 50%. It is falsified if that share is at most 20%.

4.1 Primary decision

All 41 included papers meet the primary filter (`applicable`$\geq$4). Of these, 39 fail at least two traps. That is $39/41 = 95.1\%$ `ge2`. Under the $> 50\%$ rule that share exceeds the pre-registered 50% threshold.

The two papers with `ge2`$= 0$ are both self-mod integrity audits rather than standard gain-reporting SI systems. They are Phantom Guardrails [Wang et al., 2026b] and Auditing Harness Tampering [Wang et al., 2026c]. No family falls below the majority threshold on locked coding (Table 3).

4.2 Per-trap fail rates

Table 2 lists fail rates among applicable (non-`NA`) cells. Figure 1 summarises per-trap fail rates and family `ge2` shares. Variance fails most often ($35/41 = 85.4\%$). Held-out, gate, and budget each fail on $26/41 = 63.4\%$ of papers. Length control is applicable on 28 papers and fails on all of them ($28/28 = 100\%$). We treat that 100% rate as a caveat. Many SI papers claim persistent text artifacts (memory, playbooks, cheatsheets) without a length-matched or irrelevant-text control. Thirteen papers are `NA` for length because the claimed mechanism is code, harness, or weight update without a persistent text buffer. The paper$\times$trap cell matrix is in Appendix A.

Table 2: Locked trap fail rates among applicable cells.

Trap	Fails	Applicable	Fail rate
held_out	26	41	63.4%
gate	26	41	63.4%
length_ctrl	28	28	100%
budget	26	41	63.4%
variance	35	41	85.4%

Table 3: Locked family counts and ge2 share.

Family	n	ge2	Share
self-mod	15	13	86.7%
memory	11	11	100%
prompt-opt	6	6	100%
ai-scientist	4	4	100%
reflection-scaffold	4	4	100%
other	1	1	100%
All	41	39	95.1%

4.3 Family crosstab

Table 3 breaks ge2 by family on the locked sheet. Memory, prompt-opt, ai-scientist, and reflection-scaffold are all 100% ge2 in this corpus. Self-mod is $13/15 = 86.7\%$. The two ge2$= 0$ papers sit in self-mod.

4.4 Coder agreement and lock

Pass 1 and pass 2 agree on 82% of trap cells $(168/205)$. A blind human spot-check on eight papers (40 cells) agrees with the adjudicated sheet on $35/40$ cells. Lionel accepted five flips and froze the locked sheet. Material ge2 change from those flips is limited. One accepted flip moves Harness Tampering [Wang et al., 2026c] to ge2$= 0$ via a budget pass. The locked budget cell for that paper supersedes a stale evidence note that still mentions unmatched budget.

4.5 Sensitivity analyses

Length control is the only trap with a 100% applicable fail rate. That invites the worry that ge2 is driven by a single harsh item. Table 4 reports pre-registered-style sensitivities on the locked sheet. Ignoring length yields $38/41 = 92.7\%$ ge2 on the locked sheet (pass 1 was $36/41 \approx 87.8\%$). Year slices stay above 50%. N/A-heavy papers (applicable$=$ 4, length NA) are $11/13 = 84.6\%$ ge2. Papers with all five traps applicable are $28/28 = 100\%$. A charitable full recode sheet is not available. No locked charitable coding exists. The nearest pre-registered harsh-item ablation is ignore-length. All reported sensitivities remain above 50%. The claim does not depend on counting length alone.

4.6 What the two clean papers show

Phantom Guardrails [Wang et al., 2026b] and Harness Tampering [Wang et al., 2026c] are integrity audits. They do not claim large task gains over unmatched baselines in the style of typical SI system papers. Their presence in the inclusion set is intentional. They stress-test whether the checklist only fires on weak reporting. Even with them included, majority multi-trap failure remains.

5 Limitations

Methods audit, not mechanism proof. A trap fail means the published protocol does not report the control. It does not prove that the system's gains are null under that control. Budget-matched re-runs could still leave residual SI gains on some tasks.

Length control is harsh. Every applicable length cell fails. That rate is a caveat. Ignore-length sensitivity still yields a large ge2 share, but the trap definition may over-call papers whose text artifacts are short or whose main claim is not "extra context helps."

Table 4: Locked sensitivity checks. Charitable full recode is unavailable.

Slice	ge2	Share
Ignore length (locked)	38/41	92.7%
Ignore length (pass 1)	36/41	≈87.8%
2024	9/9	100%
2025-only	19/19	100%
2026	11/13	84.6%
2025+	30/32	93.8%
Applicable = 4 (N/A-heavy)	11/13	84.6%
Applicable = 5	28/28	100%
Charitable full recode	—	not available
Primary (locked)	39/41	95.1%

Inclusion and family labels. The 41-paper freeze is judgment-laden. Family tags are mutually exclusive by construction and may hide hybrids. Integrity audits that meet inclusion stay in the SI corpus as stress tests. Adjacent benchmark and budget-control methods papers stay in Related Work because they do not claim a new SI system.

Coding subjectivity. Cell agreement is 82%, not perfect. Spot-check agreement is $35/40$. Five human flips entered the lock. Thirty-three papers were not human-reread cell by cell. Other labs could disagree on edge cases such as model-judge gates or charitable `NA` for code-only self-mod. No locked charitable full-recode sheet exists.

No new bake-off. We do not re-run CER-style systems against equal-token sampling [Liu et al., 2025]. We do not claim that every SI family collapses under matched compute. The freeze shows majority multi-trap failure in this 2024–2026 inclusion set.

Time slice. The corpus is 2024–2026 peer-visible arXiv work meeting our rule. Later papers may improve reporting without changing this snapshot.

6 Conclusion

We audited 41 frozen self-improving LLM-agent papers with five written eval traps. Primary `ge2` is $39/41 = 95.1\%$. That share exceeds the pre-registered 50% threshold. Variance, held-out, gate, and budget fails are common. Length control fails on every applicable paper and remains a caveat, yet locked ignore-length `ge2` stays at $38/41 = 92.7\%$.

Equal-token reflection studies already settled that budget mismatch can invent scaffold wins [Wang et al., 2024, Mirzaei, 2026]. Our result extends the evaluation threat story to the broader SI-agent literature as a paper population. Future SI claims should treat held-out splits, oracle-free gates, length-matched text controls, matched budgets, and multi-seed reporting as default rather than optional polish.

A Locked coding matrix

Table 5 lists every locked trap cell. Source is `coding-locked.tsv`. Readers can re-threshold without re-reading the papers. Evidence notes are omitted for length. For Harness Tampering (`2609.00069`), the locked budget cell is `pass` after a spot-check flip; that cell supersedes a stale evidence note that still mentions unmatched budget.

Table 5: Locked coding matrix. Source: `coding-locked.tsv`. Sorted by family then arXiv id.

arXiv id	family	held	gate	len	budg	var	fails	ge2
2408.06292	ai-scientist	F	F	F	F	F	5	1
2504.08066	ai-scientist	F	P	F	F	F	4	1
2511.02824	ai-scientist	F	P	F	F	F	4	1
2603.08127	ai-scientist	F	F	F	F	F	5	1
2407.04363	memory	F	P	F	F	F	4	1

Continued on next page

arXiv id	family	held	gate	len	budg	var	fails	ge2
2502.12110	memory	F	P	F	F	F	4	1
2504.07952	memory	F	F	F	P	F	4	1
2505.17716	memory	F	P	F	F	F	4	1
2506.06698	memory	F	P	F	P	F	3	1
2508.06433	memory	F	P	F	F	F	4	1
2508.16153	memory	P	P	F	F	F	3	1
2510.04618	memory	F	P	F	P	F	3	1
2510.12635	memory	F	F	F	F	F	5	1
2608.00017	memory	F	F	F	P	F	4	1
2608.18066	memory	F	P	F	F	F	4	1
2510.07841	other	F	F	NA	F	P	3	1
2406.07496	prompt-opt	P	F	F	F	P	3	1
2406.11695	prompt-opt	P	F	F	P	P	2	1
2507.19457	prompt-opt	P	F	F	P	F	3	1
2510.07475	prompt-opt	P	F	F	F	F	4	1
2604.04247	prompt-opt	P	F	F	F	F	4	1
2606.04465	prompt-opt	P	F	F	F	P	3	1
2407.18219	reflection-scaffold	P	F	NA	P	F	2	1
2502.04780	reflection-scaffold	F	F	F	F	F	5	1
2603.24639	reflection-scaffold	P	F	F	P	F	3	1
2606.31270	reflection-scaffold	F	F	NA	F	F	4	1
2404.11964	self-mod	F	P	NA	F	F	3	1
2408.08435	self-mod	P	F	F	F	P	3	1
2410.04444	self-mod	P	F	NA	F	F	3	1
2410.06153	self-mod	F	F	F	P	F	4	1
2504.15228	self-mod	F	F	NA	F	F	4	1
2505.21496	self-mod	P	F	NA	F	F	3	1
2505.22954	self-mod	F	F	NA	F	F	4	1
2508.00271	self-mod	F	P	F	F	F	4	1
2510.21614	self-mod	P	F	NA	P	F	2	1
2605.09998	self-mod	F	F	F	F	F	5	1
2607.13083	self-mod	P	P	NA	P	P	0	0
2607.21461	self-mod	F	P	F	P	F	3	1
2607.24300	self-mod	F	F	NA	P	F	3	1
2608.07645	self-mod	F	F	NA	P	F	3	1
2609.00069	self-mod	P	P	NA	P	F	1	0

Abbreviations: P = pass, F = fail, NA = not applicable.

References

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, et al. GEPA: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2507.19457.

Bjarni Haukur Bjarnason, André Silva, and Martin Monperrus. On randomness in agentic evals. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2602.07150.

Dongming Jiang, Yi Li, Songtao Wei, Jinxin Yang, Ayushi Kishore, Alysa Zhao, Dingyi Kang, Xu Hu, Feng Chen, Qiannan Li, and Bingzhe Li. Anatomy of agentic memory: Taxonomy and empirical analysis of evaluation and system limitations. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2602.19320.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. DSPy: Compiling declarative language model calls into self-improving pipelines. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.03714. ICLR 2024.

Yitao Liu, Chenglei Si, Karthik R Narasimhan, and Shunyu Yao. Contextual experience replay for self-improvement of language agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14179–14198, 2025. doi:10.18653/v1/2025.acl-long.694. URL https://arxiv.org/abs/2506.06698. ACL 2025.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2408.06292.

Iliya Mirzaei. Sample more, reflect less: Self-refine and reflexion lose to repeated sampling at equal token cost, from 1.5b to 7b. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2607.28576.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.03314. ICLR 2025.

Michelle Vaccaro. Preregistration for experiments with AI agents. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2606.11217.

Hao Wang, Hanchen Li, Qiuyang Mang, Alvin Cheung, Koushik Sen, and Dawn Song. Do androids dream of breaking the game? systematically auditing AI agent benchmarks with benchjack. *arXiv preprint*, 2026a. URL https://arxiv.org/abs/2605.12673.

Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. Reasoning in token economies: Budget-aware evaluation of LLM reasoning strategies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19916–19939, 2024. doi:10.18653/v1/2024.emnlp-main.1112. URL https://arxiv.org/abs/2406.06461. EMNLP 2024.

Su Wang, Pin Qian, Yifan Lin, Jingzhou Xu, Yihang Chen, Xiaochong Jiang, Lifei Liu, and Haoran Yu. Phantom guardrails: When self-improving agent harnesses fix failures that never happened. *arXiv preprint*, 2026b. URL https://arxiv.org/abs/2607.13083.

Xing Wang, Xiaoyi Zhang, and Jie Shao. Auditing harness tampering in self-improving agents. *arXiv preprint*, 2026c. URL https://arxiv.org/abs/2609.00069.

Yike Wang, Huaisheng Zhu, Zhengyu Hu, Yige Yuan, Zhengyu Chen, Shakti Senthil, and Hannaneh Hajishirzi. Rethinking the evaluation of harness evolution for agents. *arXiv preprint*, 2026d. URL https://arxiv.org/abs/2607.12227.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2504.08066.

Qinyuan Ye, Yu Li, Yada Pruksachatkun, Jiaxin Zhang, and Chien-Sheng Wu. On the fragility of self-improving agents: Variance, task order, and underspecification. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2608.18066.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative AI by backpropagating language model feedback. *Nature*, 639:609–616, 2025. doi:10.1038/s41586-025-08661-4. URL https://arxiv.org/abs/2406.07496. Published as Nature article; earlier arXiv title: TextGrad: Automatic "Differentiation" via Text.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents. *arXiv preprint*, 2025a. URL https://arxiv.org/abs/2505.22954.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, et al. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv preprint*, 2025b. URL https://arxiv.org/abs/2510.04618.

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Antony Kellermann, Jasjeet Sekhon, Jacob Steinhardt, Sarah Schwettmann, Arvind Narayanan, Matei A Zaharia, Ion Stoica, Percy Liang, and Daniel Kang. Establishing best practices in building rigorous agentic benchmarks. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2025. URL https://arxiv.org/abs/2507.02825. NeurIPS 2025 Datasets & Benchmarks.