

Trace Evolution Bench: Measuring Whether Agents Improve from Their Own Environment Traces

Fangfei Li,¹ Chenyang Zhao,¹ Long Wang,¹ Feng Tian,¹ Lv Guo^{1,*}

¹MatrixOrigin, China

*Corresponding author.

lifangfei@matrixorigin.cn, zhaochenyang@matrixorigin.cn,
wanglong@matrixorigin.io, tianfeng@matrixorigin.cn,
guolv@matrixorigin.cn*

Abstract

Self-evolving agents are often judged by endpoint Gain (final held-out success minus round-0 success), which can conflate true evolution with frozen rescoring variance. We present Trace Evolution Bench (TEB), an evaluation protocol that isolates attributable improvement through four controls: locked held-out identifiers, validation-gated accept/rollback, repeated frozen `no_evolve` rescoring, and seed-matched paired excess against each seed’s own noise band. TEB includes three frozen 80/20/40 packs, with coding as the primary scoreboard because success is determined by hidden unit tests in a closed workspace. On coding (40 held-out tasks), DeepSeek v4 Pro under `no_evolve` keeps version M_0 yet shifts by 7.5, 5.0, and 12.5 percentage points across seeds, showing that endpoint Gain alone overstates progress. Under this control, seven of nine DeepSeek text-update cells exceed their seed bands, but the pattern does not reproduce on GLM and MiniMax; no method remains above its own band on every seed of all three models. On separate packs, office DeepSeek does not exceed its seed bands (widest 17.5 points), and manufacturing has no paired positive excess beyond its seed band. These results position TEB as a reproducible attribution contract for self-evolving agents.

1 Introduction

Most agent benchmarks freeze a policy and score static capability [Liu et al., 2024, Jimenez et al., 2024, Yao et al., 2023, Zhou et al., 2024, Mialon et al., 2024, Shridhar et al., 2021]. Self-evolving systems pose a different measurement problem: whether post-update improvement is attributable to the evolution procedure rather than to evaluation noise or protocol mismatch [Silver et al., 2017]. In practice, attribution is confounded when teacher traces substitute for self-generated traces [Ross et al., 2011], when evaluation targets a checkpoint different from the artifact returned by `evolve()`, when held-out ids leak into the update [Magar and Schwartz, 2022], or when text updates and weight updates are compared under one budget axis [Henderson et al., 2018]. Reporting Success % alone can hide weak round-0 baselines [Liang et al., 2023]; reporting Gain from one stochastic round-0 pass can credit frozen rescoring variance as evolution [Dodge et al., 2019]. These issues motivate an attribution-first contract, not only a final-score report.

Trace Evolution Bench (TEB) operationalizes this contract with locked held-out ids, validation-gated accept/rollback, frozen `no_evolve` rescoring, and seed-matched paired excess. We report coding as the primary scoreboard because success is determined by hidden unit tests in a closed workspace: tests, rather than phrasing, decide each pass. This setting is harder to game than office’s rule-bounded delivery and broader than manufacturing table QA. The three scored methods were fixed before held-out reading, via an office validation screen whose mean gains were not positive (Section 4). On coding, DeepSeek v4

Pro [DeepSeek-AI, 2026] `no_evolve` keeps version M_0 yet moves 7.5, 5.0, and 12.5 percentage points across four passes of the same 40 held-out tasks (Table 6), showing that endpoint Gain still contains frozen rescore. Under paired excess, seven of nine DeepSeek text-update cells lie above their own seed bands, while `gepa_llm` seed 44 moves down by the width of its band (Table 8, Figure 4). GLM and MiniMax do not reproduce that seven-of-nine pattern (Figure 6, Table 9); no method stays above its own band on every seed of all three models. GLM has two paired cells above their own bands, and MiniMax `expel_llm` seed 44 is a third, with final regression 0.30. Office and manufacturing are interpreted as separate control packs on their own bands: office DeepSeek does not exceed its bands (widest 17.5 pp; Section 6), and manufacturing has no paired positive excess beyond its seed band (Section 5).

Contributions. (1) An attribution-first evaluation contract for self-evolving agents: locked held-out ids, validation-gated accept/rollback, frozen `no_evolve` rescoring, and separate budget classes for B0 text updates versus B2 weight updates. (2) A reproducible diagnostic layer for gain claims: per-seed frozen noise bands, seed-matched paired excess, round-0 pairing checks, and explicit regression/eligibility markers that separate scoring evidence from ranking policy. (3) Cross-model evidence on the coding primary scoreboard: DeepSeek shows seven of nine cells above their own bands, but this pattern does not reproduce on GLM and MiniMax; no method stays above its own band on every seed of all three models. (4) Three frozen 80/20/40 packs with separate interpretation lanes: coding as primary evidence, office and manufacturing as independent controls on their own seed bands.

2 Related work

ReAct-style loops [Yao et al., 2023] and agent suites [Liu et al., 2024, Jimenez et al., 2024] grade a frozen policy. Reflexion [Shinn et al., 2023], ExpeL [Zhao et al., 2024], self-refinement [Madaan et al., 2023], workflow memory [Wang et al., 2025], prompt evolution [Agrawal et al., 2026], skill libraries [Wang et al., 2023], and open-ended mutation [Zhang et al., 2026a] are systems TEB can host. The adapters are method-inspired text updates on one frozen actor, and the same three methods are scored on coding, office, and manufacturing. The quantity compared with those systems is whether an update beats a frozen rescore, not their published scores.

Three protocol choices separate TEB from that set. The held-out ids never enter `evolve()`. Neighbors that mix the test ids into the update, or that do not say which split is locked, are not the same contract [Magar and Schwartz, 2022]. A validation gate accepts a proposal only when validation success does not fall. Held-out success is scored after that decision, on the accepted artifact, and is not the accept test. Weight updates are a separate budget class. A B0 text update is not ranked against a B2 weight update, including LoRA-style training [Hu et al., 2022].

Table 1 is the evaluation question.

Table 2 is the feature contrast under binary coding, using a stated-evidence review. ✓ marks explicit support in the paper’s main protocol, and × marks all other cases (including not stated). Nguyen et al. [2026] commit a candidate only when held-out performance does not regress, so the validation-gate and held-out-lock cells are crosses. Zheng et al. [2026] separate training, validation, and held-out views but do not state which split accepts an update. Shi et al. [2026] separate exploration from strict held-out evaluation and compare history, summary-memory, and training pathways in one framework. The freeze-weights row is a cross where the benchmark trains: Meng et al. [2026], Kang et al. [2026], Yang et al. [2026], Yuan et al. [2026], Rank et al. [2026], and Shi et al. [2026]. PostTrainBench also forbids training on the test set and scores the submitted checkpoint. Those papers also discuss transfer, business tasks, frozen artifacts, retained experience, and whether one financial case helps the next; none of those properties is a row of Table 2 [Gao et al., 2026, Zhou et al., 2026, Lin et al., 2026, Xue et al., 2026, Deng et al., 2026]. Zheng et al. [2025] and

Table 1: Evaluation target of each benchmark. A frozen rescore band is the range of repeated held-out passes of one unchanged policy.

Benchmark	Scored unit	Frozen rescore band
SWE-bench, AgentBench [Jimenez et al., 2024, Liu et al., 2024]	One episode	No
LifelongAgentBench [Zheng et al., 2025]	A task sequence	No
SEA-Eval [Jiang et al., 2026]	A task sequence	No
EvoAgentBench [Gao et al., 2026]	Transfer on an ability-supported split	No; one vanilla condition
EvoPathBench [Lin et al., 2026]	Capability checkpoints over ordered streams	No
S ³ Gym [Shi et al., 2026]	Strict held-out evaluation	No
TEB	The same locked held-out ids	Yes; four <code>no_evolve</code> passes

Jiang et al. [2026] evaluate agents over task sequences rather than single episodes; Liu et al. [2026] studies skill feedback over rounds, and Shang et al. [2026] gates skill updates before commit.

Table 2: Protocol mechanics contrast (binary coding). ✓ means the property is explicitly supported in the paper’s main protocol; × means otherwise (including not stated). In “Carrier flex. in this paper (m/p/c/w)”, m/p/c/w = memory/prompt/code/weight. Abbreviations: SEA (SEAGym), Evo (EvoAgentBench), GDP (GDPevo), RSI (RSIBench-Data), PTB (PostTrainBench), TR (TRACE), MLE (OpenMLE), EPB (EvoPathBench), S3 (S³Gym), SEB (SE-Bench), PAST (PAST-Bench), Fin (FinEvo-Bench).

Property	TEB	RSEA	SEA	Evo	GDP	RSI	PTB	TR	MLE	EPB	S3	SEB	PAST	Fin
Own traces	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	×	✓	✓
Multi-round	✓	✓	✓	×	×	✓	×	×	×	✓	✓	×	✓	✓
Carrier flex. in this paper (m/p/c/w)	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Explore/eval split	✓	×	×	✓	×	×	✓	×	×	✓	✓	×	×	×
Validation gate	✓	×	×	×	×	×	×	×	×	×	×	×	×	×
Held-out lock	✓	×	×	✓	×	×	✓	×	×	✓	✓	×	×	×
Weight freeze	✓	✓	✓	✓	×	×	×	×	×	✓	×	×	✓	✓
Weight slot	✓	×	×	×	×	×	×	×	×	×	×	×	×	×

3 Protocol

Figure 1 is the measurement. Figure 2 is one round of it. Round 0 scores M_0 on held-out and freezes Gain’s baseline. Each later round collects a disjoint slice of evolution episodes, calls `evolve()` once, accepts the proposal iff validation success is at least the current agent’s (`min_delta=0`), otherwise rolls back, then scores the accepted agent on the same held-out ids. Held-out success never enters `evolve()`.

Let $\text{Succ}(M) = \#\{\text{held-out pass}\}/n$ be a rate on the locked held-out ids. Percentage points are 100 times a rate difference:

$$\text{Gain}(M_r) = 100 \times (\text{Succ}(M_r) - \text{Succ}(M_0)), \quad (1)$$

with this-run M_0 . In these equations M_0 is that round-0 score. A version label $M_0, M_1, \dots$ names the accepted artifact, and `no_evolve` keeps version M_0 while the score still moves. A proposal is accepted when validation success is at least the current agent’s (`min_delta = 0`); otherwise it is rolled back. Held-out ids do not enter `evolve()`. For actor a , pack e , and seed s , the noise band is the range of the four frozen

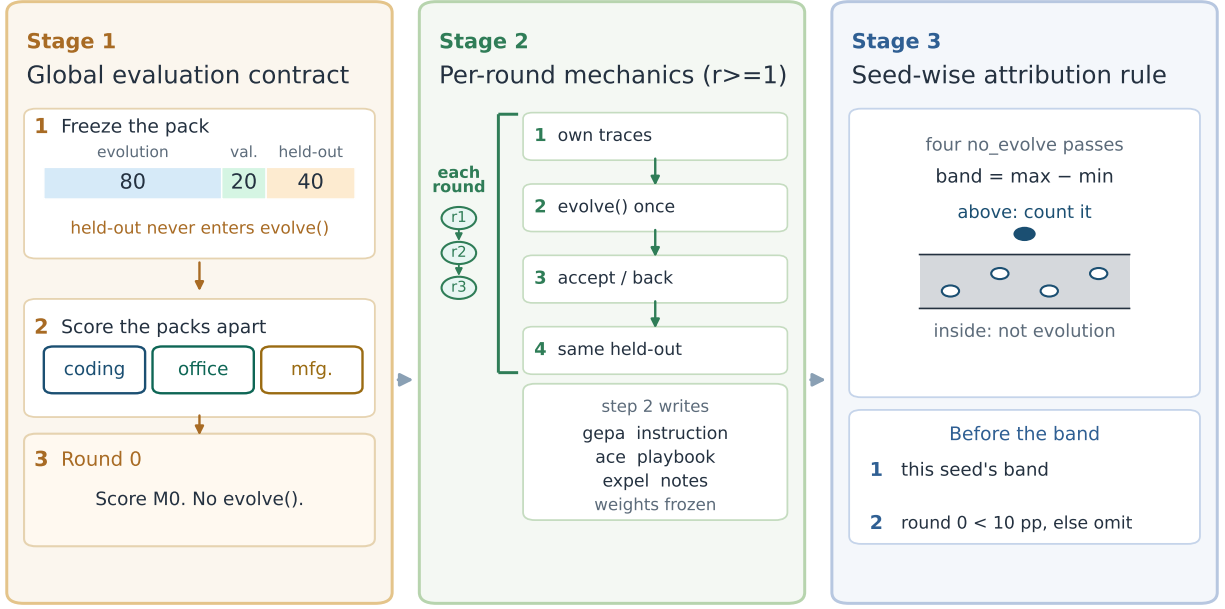


Figure 1: TEB global evaluation contract across rounds. Round 0 scores M_0 and does not update. Rounds 1–3 each accept or roll back one text update, then rescore the same held-out. Excess is read against that seed’s own four-pass band.

no_evolve held-out rates $k \in \{0, 1, 2, 3\}$:

$$B(a, e, s) = 100 \times \left(\max_k \text{Succ}(M_{a,e,s,k}^{(0)}) - \min_k \text{Succ}(M_{a,e,s,k}^{(0)}) \right). \quad (2)$$

Paired excess uses the same actor, pack, and seed:

$$\text{Excess} = \text{Gain}_{\text{method}}(a, e, s) - \text{Gain}_{\text{no_evolve}}(a, e, s). \quad (3)$$

An excess inside that seed’s own $B(a, e, s)$ is not reported as evolution. A method whose round-0 held-out rate differs from the same-seed frozen round 0 by 10 pp or more is not read as the same starting draw. Each seed is compared with its own $B(a, e, s)$. Office DeepSeek’s widest seed band is 17.5 pp, and seed 43’s own band is 7.5 pp. A value inside the widest band can still sit on seed 43’s band. Manufacturing keeps its own bands. Regression is, among tasks passed on the previous held-out pass, the fraction that fail on the final pass. A rate above 0.20 is stored on the cell. The excess comparison keeps the cell either way. Coding is the primary scoreboard. The evolution-token cap is 500,000 per round, summed from per-step usage on evolution tasks only. Validation, held-out, and calls inside `evolve()` are not added. Crossing the cap skips `evolve()` for that round, marks `over_budget`, and skips later rounds when the violation is recorded before the next round.

no_evolve returns the same agent. Its four held-out passes are repeated measurements of a frozen policy [Henderson et al., 2018, Madaan et al., 2024]. The noise band is the max-minus-min of those four scores. Paired excess is a method’s Gain minus the same model and seed’s no_evolve Gain. Excess inside the band is not reported as evolution. Runner eligibility (`main_table_eligible`) is a separate flag: a frozen no_evolve cell can pass it while `agent_version` stays M_0 .

Paper-min, unless labeled otherwise: $R=3$; at least three seeds [Bouthillier et al., 2021]; at least 40 evolution ids, 10 validation, 20 held-out; M_0 inside $[0.10, 0.80]$ or the run is flagged; `human_edits=0`.

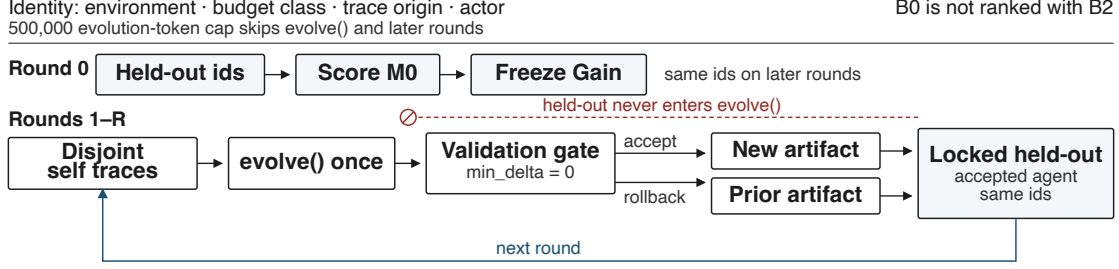


Figure 2: Per-round accept/rollback mechanics (this loop is rounds $r \geq 1$ in Figure 1). Round 0 freezes Gain on the locked held-out ids. Later rounds call `evolve()` once; accept and rollback both score that same held-out. Held-out never enters `evolve()`. B0 is not ranked against B2.

The runner also flags a pack with no OOD slice. Coding, the primary scoreboard, is one family and has no OOD slice; that flag does not move the scoreboard. Each pack uses 80/20/40, so $n=40$ held-out (2.5 pp per task). Default mix all, schedule accumulate_chunks, curriculum none [Bengio et al., 2009]. Rank only inside (env, budget_class, trace_origin, actor): B0 does not train weights and is not ranked with B2, a weight update. This paper reports B0 and B2 only; B1 is out of scope.

Table 3 is the pack contract.

Table 3: Three frozen packs. Each has 140 tasks, split 80/20/40, so held-out $n=40$. The observation omits gold files, expected SQL, and unit-test oracles.

	Office	Coding	Manufacturing
Source	Product-usage workflows	142 merged pull requests, rewritten as functions	12 real tables; 31 real questions, 109 generated
Items	10 families	One family, fix_bug	13 families
Outside evolution	12/40: access_denied, do_not_publish	None on held-out	12/40: real questions on 4 unseen table families
Tools	list_files, read_file, write_file, publish_artifact, grant_access, submit	list_dir, read_file, write_file, submit; one file, src/solution.py	list_tables, describe_table, execute_sql, submit_answer
Success	Every terminal check passes	Hidden unit tests pass	Score ≥ 80 : SQL 40, result 40, answer 20
Step cap	20	12	12

4 Coding measurement

Coding is the primary scoreboard. It uses the same 80/20/40 split, $R=3$, and B0 Standard Tool Actor. The episode cap is 12 steps. Held-out is one family, fix_bug ($n=40$). A task is one file, src/solution.py. The tools are list_dir, read_file, write_file, run_tests, and submit. Success is the hidden unit tests; the gold patch is not in the observation. Every miss is the 12-step cap, so a gain here does not open a second task type. The three methods, gepallm, aceallm, and expellm, were fixed on an office validation screen before held-out was read (Appendix A.3). That screen is a separate smoke split: one round, six tasks, ten steps. Five methods passed the artifact check. The screen kept the three with the highest mean validation gain, then lower screen regression, then fewer selection tokens. Their mean screen gains were 0, -5.6 , and -11.1 pp, none positive (Table 4). Those gains do not predict the held-out excess

on this board. Each method reads that round’s own traces and injects the text in Table 5 into the same frozen model.

Table 4: Screen ranking across three models at seed 42. Mean gain is in percentage points. Mean reg. is a rate. Valid is the artifact check.

Method	Valid	Gain (pp)	Mean reg.	Tokens
gepa_llm	true	0.0	0.056	469231
ace_llm	true	−5.6	0.111	513033
expel_llm	true	−11.1	0.056	448950
dgm	true	−16.7	0.111	396406
clin_llm	true	−16.7	0.222	411690
awm_llm	false	0.0	0.000	447296
reasoning_bank	false	−11.1	0.056	457843

Table 5: Text injected by the three scored methods. Each reads that round’s own traces. The kept text is the TEB mechanism.

Method	Injected text
gepa_llm	One revised instruction. The validation gate accepts it or rolls it back.
ace_llm	A grow-and-refine playbook from a generator, reflector, and curator.
expel_llm	A deduplicated pool of insights from success and failure.

DeepSeek `no_evolve` keeps version M_0 with four-pass ranges $B = 7.5, 5.0$, and 12.5 pp (Table 6, Figure 3). Frozen Gain and band for all three models are Table 7.

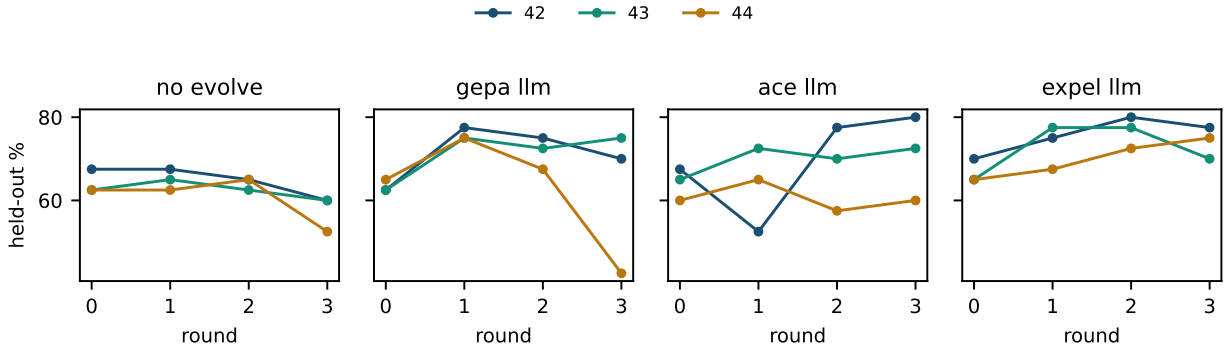


Figure 3: DeepSeek coding held-out success by round. `no_evolve` keeps version M_0 and the score still moves. Seeds are 42, 43, and 44.

Figure 3 is why excess, not the raw curve, is the coding result. `no_evolve` keeps version M_0 while each seed’s curve still moves, and its Gain is negative on all three seeds. A method panel can rise while part of that rise is the frozen score falling.

Table 7 is the cross-model control. DeepSeek is the only model whose three frozen Gains are all negative, so a rising method panel on that model includes the frozen score falling. GLM seeds 42 and 44 are also negative. GLM seed 44’s band is 15.0 pp, so an excess of $+15.0$ equals that band. The same $+15.0$ on DeepSeek seed 42 exceeds its 7.5 pp band.

Table 6: Coding, DeepSeek v4 Pro, `no_evolve` only. Success is held-out percent. Version is M_0 at every round. Range is that seed’s band.

Seed	r0	r1	r2	r3	Gain	Range
42	67.5	67.5	65.0	60.0	−7.5	7.5
43	62.5	65.0	62.5	60.0	−2.5	5.0
44	62.5	62.5	65.0	52.5	−10.0	12.5

Table 7: Coding `no_evolve` by model. Frozen Gain is the four-pass Gain. Band is that seed’s four-pass range. DeepSeek pass scores are Table 6.

Model	Seed	Frozen Gain	Band
DeepSeek	42	−7.5	7.5
DeepSeek	43	−2.5	5.0
DeepSeek	44	−10.0	12.5
GLM	42	−2.5	12.5
GLM	43	+7.5	7.5
GLM	44	−7.5	15.0
MiniMax	42	+5.0	10.0
MiniMax	43	0.0	2.5
MiniMax	44	0.0	12.5

Table 8 is DeepSeek paired excess. Seven of the nine cells are positive and larger than that seed’s own B . All seven have final regression at or below 0.20: `gepa_llm` seeds 42 and 43 (+15.0 against bands 7.5 and 5.0), `ace_llm` seeds 42 and 43 (+20.0 and +10.0), and `expel_llm` seeds 42, 43, and 44 (+15.0, +7.5, +20.0). Round-0 success on these seven cells is within 5 pp of the same-seed frozen round 0. `gepa_llm` seed 44 moves the other way (−12.5, equal to its 12.5 band) and its final regression is 0.48. The cell stays in the excess comparison. One seed of the same method moves down by the width of its band. The seven cells above their bands leave version M_0 . `ace_llm` seed 44, the positive cell inside its band, keeps version M_0 . Leaving version M_0 is not the excess: `gepa_llm` seed 44 also finishes at version M_2 , and its excess is −12.5.

Table 8: Coding-v1, DeepSeek v4 Pro. Excess is method Gain minus the same-seed `no_evolve` Gain, in percentage points. Band is that seed’s four-pass range. Reg. is the final regression rate. None of these nine cells is `over_budget`.

Method	Seed	Gain	Frozen	Excess	Band	Reg.	Against that band
gepa	42	+7.5	−7.5	+15.0	7.5	0.167	outside
gepa	43	+12.5	−2.5	+15.0	5.0	0.069	outside
gepa	44	−22.5	−10.0	−12.5	12.5	0.481	equal, down
ace	42	+12.5	−7.5	+20.0	7.5	0.065	outside
ace	43	+7.5	−2.5	+10.0	5.0	0.071	outside
ace	44	0.0	−10.0	+10.0	12.5	0.130	inside
expel	42	+7.5	−7.5	+15.0	7.5	0.062	outside
expel	43	+5.0	−2.5	+7.5	5.0	0.097	outside
expel	44	+10.0	−10.0	+20.0	12.5	0.103	outside

GLM coding is the same comparison on a wider frozen range (Table 9). `gepa_llm` excesses are +7.5,

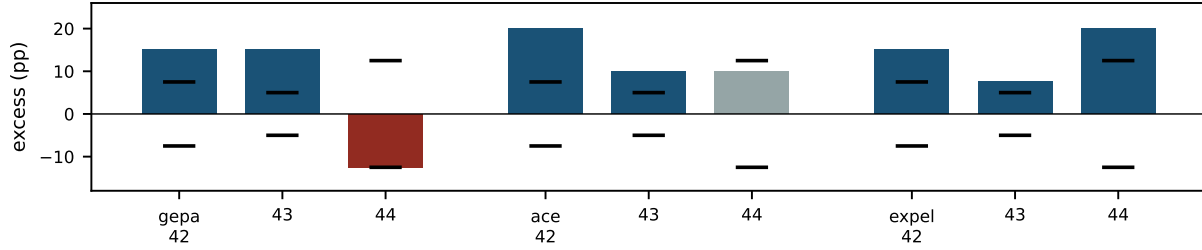


Figure 4: DeepSeek coding paired excess. Black ticks are that seed’s own band. Dark bars lie strictly above it. The light positive bar is `ace_11m` seed 44, inside its 12.5 pp band. `gepa_11m` seed 44 equals its band on the way down.

0, and +15.0 against bands 12.5, 7.5, and 15.0: the first two sit inside, and seed 44 equals its band. `ace_11m` seed 42 is -12.5 on a 12.5 band and is `over_budget`; seeds 43 and 44 are +5.0 (inside 7.5) and +15.0 (equal to 15.0), and seed 43 is also `over_budget`. `expel_11m` seeds 43 and 44 are +10.0 and +22.5 against bands 7.5 and 15.0, with final regression 0.03 and 0.13. Seed 44’s round 0 is 60.0 against the frozen round 0 of 67.5, a 7.5 pp gap, below the 10 pp cutoff, so that excess stays in the comparison. `expel_11m` seed 42 is not in this pairing: its recorded M_0 is 5%, while `no_evolve` seed 42 opens at 72.5%. The round-0 gap is 67.5 pp. The cell is outside the paper-min M_0 window [0.10, 0.80].

MiniMax coding is finished for all three methods on seeds 42, 43, and 44. On seed 42 the excesses are -7.5, -5.0, and +7.5 for `gepa_11m`, `ace_11m`, and `expel_11m`, all inside the 10.0 band; `expel_11m` has final regression 0.22. Seed 43 `gepa_11m` and `ace_11m` open at 27.5% against a frozen round 0 of 37.5%, so their excesses of +45.0 and +17.5 are not the same starting draw. `expel_11m` seed 43 equals the 2.5 band and has regression 0.36. On seed 44, `gepa_11m` Gain is +10.0 and its excess is +10.0, inside the 12.5 band, with regression 0.14; `ace_11m` excess is 0. `expel_11m` excess is +15.0, above that 12.5 band. Its round 0 is 37.5 against the frozen round 0 of 40.0, and its final regression is 0.30.

Table 9: Coding status for GLM and MiniMax. Excess and band are defined as in Table 8. Unpaired means the round-0 gap is at least 10 pp. Reg. is the final regression rate. Cap marks `over_budget`. A regression above 0.20 is shown and does not remove the cell.

Model	Method	Seed	Excess	Band	Reg.	Cap	Against that band
GLM	gepa	42	+7.5	12.5	0.207		inside
GLM	gepa	43	0.0	7.5	0.069		inside
GLM	gepa	44	+15.0	15.0	0.094		equal
GLM	ace	42	-12.5	12.5	0.130	yes	equal, down
GLM	ace	43	+5.0	7.5	0.065	yes	inside
GLM	ace	44	+15.0	15.0	0.091		equal
GLM	expel	42	+70.0	12.5	0.000		unpaired
GLM	expel	43	+10.0	7.5	0.031		outside
GLM	expel	44	+22.5	15.0	0.125		outside
MiniMax	gepa	42	-7.5	10.0	0.353		inside
MiniMax	gepa	43	+45.0	2.5	0.000		unpaired
MiniMax	gepa	44	+10.0	12.5	0.136		inside
MiniMax	ace	42	-5.0	10.0	0.214		inside
MiniMax	ace	43	+17.5	2.5	0.158		unpaired
MiniMax	ace	44	0.0	12.5	0.188		inside
MiniMax	expel	42	+7.5	10.0	0.222		inside
MiniMax	expel	43	+2.5	2.5	0.364		equal
MiniMax	expel	44	+15.0	12.5	0.304		outside

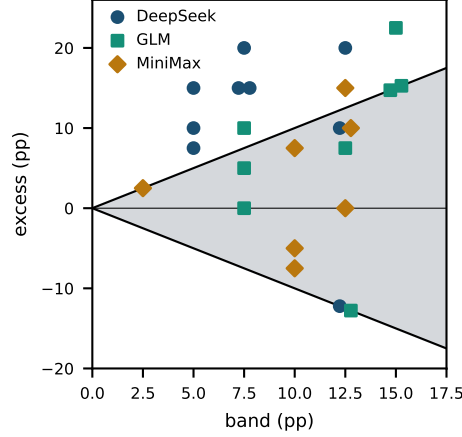


Figure 5: Paired cells. The wedge is inside the band. A point above the upper line clears that seed’s band. Unpaired cells are omitted.

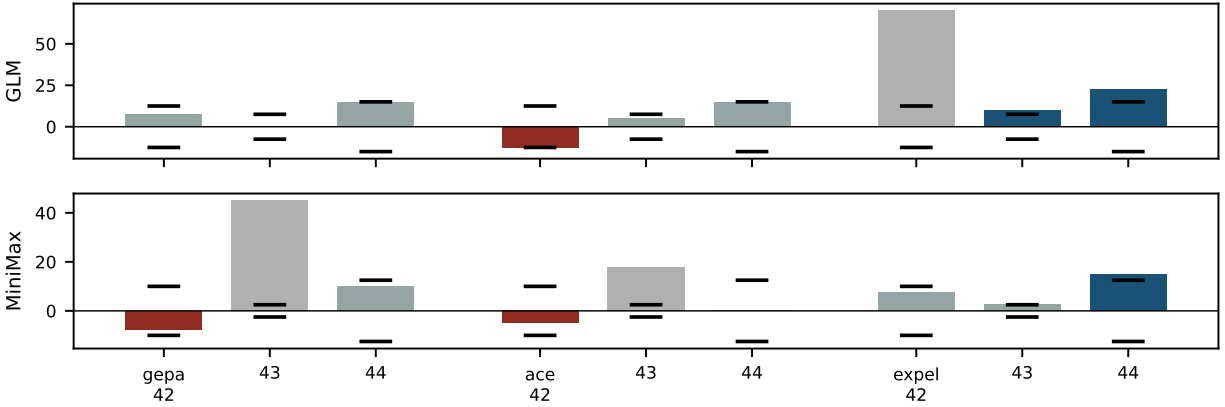


Figure 6: GLM and MiniMax coding paired excess. Black ticks are that seed’s own band. Dark bars lie strictly above it: GLM `expel_llm` seeds 43 and 44, and MiniMax `expel_llm` seed 44. Hatched bars are not the same starting draw: GLM `expel_llm` seed 42, and MiniMax `gepa_llm` and `ace_llm` seed 43.

Figure 6 is why the seven-of-nine pattern stays a DeepSeek result. Three dark bars clear their own bands, not seven of nine. A hatched bar is a different starting draw and is not one of those seven. Figure 5 is the same paired cells. DeepSeek’s seven cells lie above the upper boundary. GLM’s +15 at seed 44 lies on the boundary of a 15 pp band. DeepSeek’s +15 at seed 42 lies above the boundary of a 7.5 pp band.

5 Manufacturing measurement

Manufacturing asks whether the coding seven-of-nine pattern appears for the same three methods. It does not. A pass lists tables, describes one, executes SQL, and submits an answer. Success is a deterministic score of at least 80: SQL 40, result 40, answer 20. The expected SQL is not in the observation. The pack uses the same 80/20/40 split and the 12-step cap. DeepSeek `no_evolve` ranges are 7.5, 5.0, and 2.5 pp. GLM ranges are 2.5, 5.0, and 2.5. MiniMax ranges are 12.5, 10.0, and 12.5. One task is still 2.5 pp.

No finished DeepSeek manufacturing cell has a positive excess larger than its seed band (Table 10). `ace_llm` excess is -2.5 on every seed; Gain on those three cells is 0. `expel_llm` seed 44 is $+2.5$, equal to the 2.5 band. `gepa_llm` seeds 42 and 43 have excesses -10.0 and -12.5 . Seed 42 `gepa_llm` and seed 42 `expel_llm` both open at 45.0 against a frozen round 0 of 35.0, so those two excesses are not the same starting draw. `gepa_llm` seed 43 opens at 47.5 against 40.0, a 7.5 pp gap, and stays in the comparison. `ace_llm` seed 44 and `expel_llm` seed 43 have final regression 0.21, above 0.20. GLM’s largest positive excess is $+5.0$ (`ace_llm` seed 43), equal to that seed’s 5.0 band. `expel_llm` seed 42 is -7.5 against a 2.5 band. The other GLM excesses are 0 or $+2.5$ on bands of 2.5 or 5.0.

Table 10: Manufacturing-v1, DeepSeek v4 Pro. Excess and band are defined as in Table 8.

Method	Seed	Gain	Frozen	Excess	Band	Against that band
<code>gepa</code>	42	-7.5	$+2.5$	-10.0	7.5	outside, negative; r0 gap 10 pp
<code>gepa</code>	43	-10.0	$+2.5$	-12.5	5.0	outside, negative; r0 gap 7.5 pp
<code>gepa</code>	44	0.0	$+2.5$	-2.5	2.5	equal
<code>ace</code>	42	0.0	$+2.5$	-2.5	7.5	inside
<code>ace</code>	43	0.0	$+2.5$	-2.5	5.0	inside
<code>ace</code>	44	0.0	$+2.5$	-2.5	2.5	equal; reg. 0.21
<code>expel</code>	42	-7.5	$+2.5$	-10.0	7.5	outside, negative; r0 gap 10 pp
<code>expel</code>	43	$+2.5$	$+2.5$	0.0	5.0	inside; reg. 0.21
<code>expel</code>	44	$+5.0$	$+2.5$	$+2.5$	2.5	equal

MiniMax manufacturing is finished for all three methods on seeds 42, 43, and 44. On seed 42, `ace_llm` excess is -15.0 and `expel_llm` excess is -7.5 against a 12.5 band, so only `ace_llm` falls outside, and it falls below. `gepa_llm` opens at 27.5 against a frozen round 0 of 17.5, a 10 pp gap, so its -7.5 excess is not the same starting draw. Seed 43 `gepa_llm` excess is $+5.0$ and `expel_llm` excess is -5.0 , both inside the 10.0 band. `ace_llm` seed 43 excess is $+20.0$ against that band, but round 0 is 12.5 against a frozen round 0 of 22.5. On seed 44, `ace_llm` excess is $+2.5$ and `expel_llm` excess is -7.5 , both inside the 12.5 band; `expel_llm` keeps version M_0 . `gepa_llm` seed 44 Gain is $+25.0$ and its excess is $+20.0$ against that band, but round 0 is 12.5 against a frozen round 0 of 22.5. Final regression on all nine MiniMax method cells is above 0.20.

6 Office measurement

Office-v1 uses the same 80/20/40 split and the B0 Standard Tool Actor. The episode cap is 20 steps. The tools are `list_files`, `read_file`, `write_file`, `publish_artifact`, `grant_access`, and `submit`. Success is every terminal check passing. A failed `submit` returns the failed check names and leaves the episode open. Gold files are not in the observation. It is scored on its own band and is not averaged with coding. Actor, prompt, and held-out ids stay fixed, and the request temperature is 0. The hosted endpoint is still not deterministic, so repeated scoring of the same version- M_0 policy is a band [Atil et al., 2024]. DeepSeek v4 Pro `no_evolve` keeps version M_0 . The three seed ranges are 17.5, 7.5, and 17.5 pp, so the widest frozen swing is 17.5 pp; one task is 2.5 points. GLM seed 42 spans 17.5 pp. MiniMax seed 42 spans 5.0 pp. GLM and MiniMax are one seed on this pack.

No DeepSeek office cell has a positive excess larger than its own seed band (Table 11). Mean excess is -2.5 pp (`gepa_llm`), -3.3 pp (`ace_llm`), and $+4.2$ pp (`expel_llm`), all inside the widest 17.5 pp range. Seed 43’s own band is 7.5 pp: `gepa_llm` is -7.5 and `ace_llm` is $+7.5$, so both sit on that band rather than beyond it. `incident_track`, `postmortem`, and `onboarding` stay at zero solved tasks on every DeepSeek pass. The solved-count changes that remain are in families the frozen policy already solves on some pass.

Table 11: Office-v1, DeepSeek v4 Pro. Excess and band are defined as in Table 8. Reg. is the final regression rate.

Method	Seed	Gain	Frozen	Excess	Band	Reg.	Against that band
gepa	42	+7.5	+10.0	-2.5	17.5	0.250	inside
gepa	43	0.0	+7.5	-7.5	7.5	0.375	equal
gepa	44	+17.5	+15.0	+2.5	17.5	0.143	inside
ace	42	+5.0	+10.0	-5.0	17.5	0.267	inside
ace	43	+15.0	+7.5	+7.5	7.5	0.077	equal
ace	44	+2.5	+15.0	-12.5	17.5	0.385	inside
expel	42	+17.5	+10.0	+7.5	17.5	0.154	inside
expel	43	+10.0	+7.5	+2.5	7.5	0.000	inside
expel	44	+17.5	+15.0	+2.5	17.5	0.091	inside

GLM excesses sit inside 17.5 pp, and the three adapter cells are over_budget. MiniMax gepa_llm excess is +2.5 pp, inside the 5.0 pp band. ace_llm excess is -7.5 pp and stopped after round 2. expel_llm excess is +7.5 pp, 2.5 pp past that 5.0 pp band, on one seed, at final version M_1 . That is not a confirmed evolution result.

7 Limitations

Coding held-out is one family, fix_bug ($n=40$, 2.5 pp per task). The seven cells above their own bands do not include gepa_llm seed 44, which moves down by 12.5 pp with final regression 0.48. GLM and MiniMax do not reproduce that seven-of-nine pattern. GLM expel_llm seeds 43 and 44 still lie above their own bands; seed 42 is unpaired (M_0 5% against a frozen round 0 of 72.5%). MiniMax expel_llm seed 44 also lies above its 12.5 band, with final regression 0.30. Frozen coding seeds 42 and 44 have regression 0.231 and 0.308 with no policy change.

Office-v1 held-out has $n=40$. Effects under the 17.5 pp DeepSeek band are unresolved by this design [Dror et al., 2018, Card et al., 2020], and the 0.20 regression cutoff already raises the runner’s flag on frozen seed 43 (0.231, M_0). Validation has $n=20$ (5 pp per task). On office-v1, MiniMax and GLM are one seed. GLM adapters and MiniMax ace_llm crossed the evolution-token cap. awm_llm, reasoning_bank, clin_llm, and dgm were screened and are not on the office scoreboard. Closed hosted models were not weight-updated; TRACE and a B2 open-weight card [Hu et al., 2022] are a different budget class.

Archived ticket-gym, inbox, and NL2SQL cards use another tool set and $n=20$ held-out and are not office results (Appendix A.18).

Manufacturing is not a coding ranking (Section 5). Office uses 20 steps per task; coding and manufacturing use 12. B0 token costs outside the evolution-token cap are estimates. MiniMax manufacturing is complete on three seeds. The two MiniMax excesses outside their seed bands (ace_llm seed 43 and gepa_llm seed 44) are not the same starting draw, and all nine MiniMax method cells have final regression above 0.20.

8 Conclusion

TEB’s main result is attribution, not a single global winner. On coding, endpoint Gain includes frozen rescoring variance: DeepSeek no_evolve keeps version M_0 while the three seeds span 7.5, 5.0, and 12.5 pp (Section 4). After seed-matched paired excess, seven of nine DeepSeek text-update cells lie above their own bands, but this pattern does not reproduce on GLM and MiniMax. Across all three models, no method

stays above its own band on every seed (Table 9).

This separation between observed Gain and attributable excess is the core contribution: TEB turns self-evolution claims into a falsifiable protocol with explicit controls, rather than a single endpoint delta. The same contract keeps office and manufacturing as independent control packs on their own bands. Office DeepSeek does not exceed its own bands (widest 17.5 pp; Section 6), and manufacturing has no paired positive excess beyond its seed band.

AI use statement

Generative AI coding assistants participated in the early execution of the implementation, namely the TEB runner and the method adapters. They also participated in polishing some prose, in retrieval, and in drafting some sections. The experimental design was set by the authors. We have not used generative AI to generate synthetic evaluation packs, formulate mathematical claims, write proofs, or invent held-out scores. All Success %, Gain, family counts, and gate decisions are read from stored run artifacts. The authors reviewed that implementation and take responsibility for the final content of this work.

Ethics statement

Manufacturing-v1 derives from internal business tables provided by a manufacturing partner. The benchmark replaces counterparty, company, department, cost-center, and document identifiers with pseudonyms and withholds the private mapping. The tables contain no personal names, phone numbers, e-mail addresses, or identity numbers. Office and coding tasks are rewritten from product workflows and pull requests; contact details in office files use reserved example domains. No human subjects were involved.

Reproducibility statement

Section 3 defines every scored quantity. Appendix A.2 lists what each adapter keeps, Appendix A.3 the development screen, Appendix A.9 the eligibility flags, and Appendix A.17 one task from each suite. The release contains the runner and table contract (`teb/`, `teb/table_contract.py`), the three packs with source metadata removed (`tasks/office`, `tasks/coding`, `tasks/manufacturing`), the configs of the reported runs, and the sanitized per-run summaries used by `scripts/plot_office_v1.py`. Appendix A.1 provides the dataset card and licenses (data CC BY 4.0, code MIT). Hosted endpoints require the reader’s own API key.

References

- Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziemis, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026.
- Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. Non-determinism of “deterministic” LLM settings. *arXiv preprint arXiv:2408.04667*, 2024.

- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, 2009.
- Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, 2021.
- Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.
- DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.
- Bo Deng, Kang Zhou, Lifan Guo, Chongyang Tao, Xuanren Chen, Chenggang Xie, Renzhao Liang, Feng Chen, and Chi Zhang. FinEvo-Bench: A longitudinal benchmark for self-evolving agents in professional financial workflows. *arXiv preprint arXiv:2608.06144*, 2026.
- Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.
- Aparna Elangovan, Jiayuan He, and Karin Verspoor. Memorization vs. generalization: Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.
- Xingze Gao, Chuanrui Hu, Hongda Chen, Pengfei Yao, Zhao Wang, Yi Bai, Zhengwei Wu, Yunyun Han, Xiaofeng Cong, Jie Gui, Yafeng Deng, and Teng Li. EvoAgentBench: Benchmarking agent self-evolution via ability transfer. *arXiv preprint arXiv:2607.05202*, 2026.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- Sihang Jiang, Lipeng Ma, Zhonghua Hong, Keyi Wang, Zhiyu Lu, Tengfei Wang, Shisong Chen, Jinghao Zhang, Tianjun Pan, Weijia Li, Jiaqing Liang, and Yanghua Xiao. SEA-Eval: A benchmark for evaluating self-evolving agents beyond episodic assessment. *arXiv preprint arXiv:2604.08988*, 2026.
- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024.

- Hangoo Kang, Tarun Suresh, Jon Saad-Falcon, and Azalia Mirhoseini. TRACE: Capability-targeted agentic training. In *International Conference on Machine Learning*, 2026.
- Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.
- Jinyang Li, Binyuan Hui, Ge Qu, Jiayi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C. C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. Can LLM already serve as a database interface? a BIRD bench for large-scale database grounded text-to-SQLs. In *Advances in Neural Information Processing Systems*, 2023.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.
- Hongqiang Lin, Chao Liu, Xiaofan Bai, Xuan Jin, Yuhong Li, Nenggan Zheng, and Xipeng Cao. Beyond endpoint performance: Process-level evaluation of self-evolving agents. *arXiv preprint arXiv:2609.24663*, 2026.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024.
- Yuxuan Liu, Zhaochen Su, Yuhao Zhang, Jiahe Guo, Zhongwei Xie, Huihao Jing, Lingyun Xie, Qing Zong, Yauwai Yim, Zhixiong Zhang, Haoran Li, and Yangqiu Song. Rethinking self-evolving agent skills: Feedback dynamics over multiple rounds. *arXiv preprint arXiv:2608.02636*, 2026.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallman, Dongchan Gao, Christopher Re, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.
- Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.
- Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 157–165, 2022.
- Bodhisattwa Prasad Majumder, Bhavana Dalvi Mishra, Peter Jansen, Oyvind Tafford, Niket Tandon, Li Zhang, Chris Callison-Burch, and Peter Clark. CLIN: A continually learning language agent for rapid task adaptation and generalization. In *COLM*, 2024.
- Fanqing Meng, Lingxiao Du, Qiguang Chen, Ziqi Zhao, Haocheng Lu, Mengkang Hu, and Michael Qizhe Shieh. RSIBench-Data: Benchmarking data-centric research for recursive self-improvement. *arXiv preprint arXiv:2607.25886*, 2026.
- Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations*, 2024.
- MiniMax. The MiniMax-M2 series: Mini activations unleashing max real-world intelligence. *arXiv preprint arXiv:2605.26494*, 2026.

- Michael Nguyen, Quoc Nguyen, and Paul Vuong. Recursive self-evolving agents via held-out selection. *arXiv preprint arXiv:2606.28374*, 2026.
- Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. ReasoningBank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.
- Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Ben Rank, Hardik Bhatnagar, Ameya Prabhu, Shira Eisenberg, Karina Nguyen, Matthias Bethge, and Maksym Andriushchenko. PostTrainBench: Can LLM agents automate LLM post-training? In *International Conference on Machine Learning*, 2026.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, 2020.
- Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 627–635, 2011.
- Linfang Shang, Ming Xu, Yiding Sun, Tianle Xia, Lingxiang Hu, Lan Xu, and Ning Zheng. When self-evolution backfires: Pre-commit gating against skill contamination in LLM agents. *arXiv preprint arXiv:2608.05810*, 2026.
- Jiajun Shi, Siyuan Tao, Yuhao Wu, Zexuan Wang, Jingyuan Zhang, et al. S3gym: Can llms turn self-testing and self-judging into self-improvement? *arXiv preprint arXiv:2608.31100*, 2026.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.
- David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550:354–359, 2017.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Aniket Mandlekar, Chaowei Xiao, Yuke Zhu, Song An, et al. Voyager: An open-ended embodied agent with large language models. In *Advances in Neural Information Processing Systems*, 2023.
- Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 63897–63911, 2025.

- Shuhan Xue, Zixin Ding, Yichen Shen, Yinjie Wang, Zhenfei Yin, Yingcheng Wu, Yuxin Chen, Mengdi Wang, and Ling Yang. PAST-Bench: Benchmarking the foundations of recursive self-improvement in personal agents. *arXiv preprint arXiv:2608.04003*, 2026.
- Junlin Yang, Che Jiang, and Kaiyan Zhang. Frontis-MA1: Training an AI4AI model towards recursive self-improvement in machine learning engineering. *arXiv preprint arXiv:2607.28568*, 2026.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. Re-Act: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.
- Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
- Jiarui Yuan, Tailin Jin, Weize Chen, and Zeyuan Liu. SE-Bench: Benchmarking self-evolution with knowledge internalization. *arXiv preprint arXiv:2602.04811*, 2026.
- Yunpeng Zhai, Shuchang Tao, Cheng Chen, Anni Zou, Ziqian Chen, Qingxu Fu, Shinji Mai, Li Yu, Jiaji Deng, Zouying Cao, Zhaoyang Liu, Bolin Ding, and Jingren Zhou. AgentEvolver: Towards efficient self-evolving agent system. *arXiv preprint arXiv:2511.10395*, 2025.
- Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel machine: Open-ended evolution of self-improving agents. In *International Conference on Learning Representations*, 2026a.
- Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, 2026b.
- 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.
- Congjie Zheng, Chuanyi Xue, Bin Liang, Jun Yang, and Changshui Zhang. SEAGym: An evaluation environment for self-evolving LLM agents. *arXiv preprint arXiv:2606.17546*, 2026.
- Junhao Zheng, Xidi Cai, Qiuke Li, Duzhen Zhang, Zhong-Zhi Li, Yingying Zhang, Le Song, and Qianli Ma. LifelongAgentBench: Evaluating LLM agents as lifelong learners. *arXiv preprint arXiv:2505.11942*, 2025.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.
- Ruiqi Zhong, Tao Yu, and Dan Klein. Semantic evaluation for text-to-SQL with distilled test suites. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Leijun Zhou, Zhihao Liu, Xiang Qu, Chenxu Liu, Yifei Liu, Yanke Yu, Jingzhe Xu, Xuejun Wu, Buyue Qian, Xi Chen, Yaowei Zheng, and Junhao Hu. GDPevo: Evaluating agent self-evolution on real business tasks. *arXiv preprint arXiv:2608.03764*, 2026.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

Appendix contents

A.1	Dataset card	18
A.2	What each office adapter keeps	19
A.3	Development screen	19
A.4	Office validation gates	20
A.5	Per-round held-out record	20
A.6	DeepSeek family counts on every pass	22
A.7	Seed-42 family counts outside DeepSeek	22
A.8	Failure reasons outside the DeepSeek office cells	24
A.9	Eligibility flags	24
A.10	Office grids	25
A.10.1	Frozen band	26
A.10.2	Paired excess	26
A.10.3	Same agent, different score	26
A.10.4	Full grid	27
A.10.5	Seed-42 transfer	28
A.10.6	Where the 40 tasks move	28
A.10.7	Failure reasons	29
A.11	Office artifact versions	29
A.12	Manufacturing frozen repeat	30
A.13	Manufacturing family counts	30
A.14	Coding frozen repeat	30
A.15	Extra office frozen passes	31
A.16	Coding grids	31
A.17	Worked examples	31
A.17.1	One task from each current suite	32
A.17.2	What <code>evolve()</code> actually returns	33
A.17.3	One held-out task, same M_0 , two scores	34
A.18	Archived ticket-gym mock	34
A.18.1	Why the protocol constraints	35
A.18.2	Success, Gain, and actors	36
A.18.3	Setup of the archived mock	36
A.18.4	Summary boards and expansion families	36
A.18.5	Round curves	36
A.18.6	Mechanism: what heuristic failures teach	37
A.18.7	Why $R=3$ is enough on ticket-gym	38
A.18.8	Inbox mock	38
A.18.9	Seed, wrapper, inbox, and hygiene tables	38

A.18.10	Weight-track provenance	38
A.18.11	Ticket-gym B0 ablations	38
A.18.12	Ticket-gym gate decisions	38
A.18.13	Ticket-gym native pairing	38
A.18.14	NL2SQL teacher contrast	39
A.18.15	Ticket-gym expansion pack	40
A.18.16	Cross-environment summary	51

A.1 Dataset card

Following datasheet practice [Gebru et al., 2021], TEB ships three frozen packs with the same 80/20/40 split and fixed held-out membership (Table 12). Each pack is adapted from a real source: product-usage workflows for office, merged pull requests of one production repository for coding, and one manufacturer’s internal business tables for manufacturing. Manufacturing contains 31 real questions; the other 109 are generated over the same real tables with recomputed gold, so “real-derived” refers to the tables and those 31 questions, not every item. The generated questions come from deterministic templates whose gold is computed by executing SQL on the released tables; a candidate is rejected if it repeats an original question, gold SQL, or answer value, names a real entity, or touches a held-out-only table. Coding tasks restate each pull request’s behavior as a rewritten Python function; no repository code is copied. Two held-out sets contain items outside the evolution distribution: office adds two held-out-only families, and manufacturing places its 12 held-out real questions on four table families that never appear in evolution or validation. A frozen DeepSeek v4 Pro policy leaves headroom on every pack, and five of the 13 manufacturing families are unsolved on every pass (Appendix A.13). Gold files, expected SQL, and unit-test oracles never enter the observation. The packs are not claimed to measure production performance or open-ended NL2SQL generalization [Yu et al., 2018, Li et al., 2023]; manufacturing scoring checks execution rather than string match [Zhong et al., 2020]. Verifiers are deterministic. The reference solution of every one of the 420 items (gold files, gold patch, or gold SQL) passes its verifier, and the released tests check this. We did not conduct a separate annotator audit. Benchmark data are licensed under CC BY 4.0 and code under MIT.

Table 12: TEB pack card. Frozen range is held-out Success over the four passes of DeepSeek v4 Pro `no_evolve`, seeds 42–44. Language counts instructions containing Chinese.

	office-v1	coding-v1	manufacturing-v1
Source	Product-usage workflows in a closed catalog workspace	142 merged pull requests of one production repository	12 pseudonymized real tables (32,290 rows) and 31 real questions
Items	140; 10 families	140; <code>fix_bug</code>	140; 13 families; 31 real, 109 generated
Held-out OOD	12/40: <code>access_denied</code> , <code>do_not_publish</code>	none	12/40: real questions on 4 unseen table families
Language	33/140 Chinese, rest English	Chinese	Chinese
Observation	Catalog files and six tools	One file, <code>src/solution.py</code>	Four tools: list tables, describe a table, execute SQL, submit an answer
Success	Every terminal check passes	Closed-workspace unit tests pass	Score ≥ 80 of 100: SQL 40, result 40, answer 20; a partial SQL or result check caps the score below 80
Step cap	20	12	12
Frozen range	22.5–45.0%	52.5–67.5%	35.0–45.0%

A.2 What each office adapter keeps

All seven adapters read the current round’s own traces, update once before the validation gate, and inject bounded text into the same frozen model. None updates weights. `no_evolve` returns the same agent and still runs the round schedule, so the baseline is round-matched. The slots are method-inspired by Expel [Zhao et al., 2024], workflow memory [Wang et al., 2025], ACE [Zhang et al., 2026b], GEPA [Agrawal et al., 2026], ReasoningBank [Ouyang et al., 2026], CLIN [Majumder et al., 2024], and DGM [Zhang et al., 2026a]. Upstream benchmark numbers are not reused.

Table 13: Method-inspired slots on the Standard Tool Actor. The kept column is the TEB mechanism. The omitted column is the upstream system this paper does not claim to reproduce.

Slot	Kept	Omitted
<code>expel_llm</code>	Insights from success and failure; deduplicated pool	Upstream harness and retrieval
<code>awm_llm</code>	Workflows from successful traces only	Upstream induction stack
<code>ace_llm</code>	Generator, reflector, curator; grow-and-refine playbook	Upstream application runtime
<code>gepa_llm</code>	One revised instruction; the TEB gate accepts or rolls it back	Population search and Pareto compile
<code>reasoning_bank</code>	Title, when, and content memories; lexical top- <i>k</i>	Upstream agent and test-time search
<code>clin_llm</code>	Condition, action, and effect rules	Domain gold memory
<code>dgm</code>	Archive, parent score, text mutation	Repository self-modification

`dgm_native_code` is a separate board (Appendix A.18.13). AgentEvolver [Zhai et al., 2025], the SEED context probe, and TRACE [Kang et al., 2026] are not part of Table 28. TRACE remains a B2 card: contrast traces, synthesize environments, train LoRA experts [Hu et al., 2022]. A closed hosted model cannot take that update.

A.3 Development screen

The screen file records `held_out_read=false`. Gain in this section is validation success after one update minus the model-matched `no_evolve` validation success on the same screen pass. Screen regression is $\max(0, \text{before} - \text{after})$ on that validation pass. It is not the held-out regression rate in Table 11. Invalid rows failed the artifact check: the evolve artifact had to carry a hash, a non-empty context state, and `implementation_fidelity=method-inspired`. The three valid methods with the highest mean gain, then lower screen regression, then fewer selection tokens, were extended to seeds 43 and 44. Those means are 0, -0.056 , and -0.111 for `gepa_llm`, `ace_llm`, and `expel_llm`: 0, -5.6 , and -11.1 pp. None is a positive validation gain. The screen budget is one round, 10 steps per task, and 6 tasks per round, with `pack_role=smoke` on a separate split. Formal office-v1 is $R=3$, 20 steps, and 80/20/40. Selection used the shorter episode. This screen is an office validation screen. The three methods it fixed are the ones scored on the coding primary scoreboard; that board was not a separate coding pre-registration (Section 4). The ranking is Table 4.

Table 14: Screen cells. Validation columns are success rates. Gain is after minus the model-matched `no_evolve` rate.

Model	Method	Before	After	Gain	Reg.	Artifact
deepseek-v4-pro	<code>expel_llm</code>	0.167	0.167	-0.333	0.000	true
deepseek-v4-pro	<code>awm_llm</code>	0.333	0.333	-0.167	0.000	true

Model	Method	Before	After	Gain	Reg.	Artifact
deepseek-v4-pro	ace_llm	0.167	0.333	-0.167	0.000	true
deepseek-v4-pro	gepa_llm	0.500	0.333	-0.167	0.167	true
deepseek-v4-pro	reasoning_bank	0.333	0.167	-0.333	0.167	false
deepseek-v4-pro	clin_llm	0.500	0.333	-0.167	0.167	true
deepseek-v4-pro	dgm	0.167	0.167	-0.333	0.000	true
minimax-m2-5	expel_llm	0.167	0.167	+0.000	0.000	true
minimax-m2-5	awm_llm	0.000	0.167	+0.000	0.000	false
minimax-m2-5	ace_llm	0.333	0.167	+0.000	0.167	true
minimax-m2-5	gepa_llm	0.000	0.167	+0.000	0.000	true
minimax-m2-5	reasoning_bank	0.000	0.000	-0.167	0.000	true
minimax-m2-5	clin_llm	0.000	0.167	+0.000	0.000	true
minimax-m2-5	dgm	0.167	0.000	-0.167	0.167	true
glm-5-3	expel_llm	0.500	0.333	+0.000	0.167	true
glm-5-3	awm_llm	0.500	0.500	+0.167	0.000	false
glm-5-3	ace_llm	0.500	0.333	+0.000	0.167	true
glm-5-3	gepa_llm	0.500	0.500	+0.167	0.000	true
glm-5-3	reasoning_bank	0.333	0.500	+0.167	0.000	true
glm-5-3	clin_llm	0.500	0.000	-0.333	0.500	true
glm-5-3	dgm	0.500	0.333	+0.000	0.167	true

A.4 Office validation gates

The gate compares validation success of the current agent with the proposal. Both numbers are one pass on $n=20$ validation tasks. A drop rolls the proposal back. Table 15 is every non-zero round in the 20-cell matrix. Round 0 has no proposal. One validation task is 5 pp. Several accepts in Table 15 are 0 pp changes, which `min_delta=0` keeps. On `no_evolve` the artifact stays M_0 whether the row says accept or rollback: the two validation numbers are two passes of one frozen policy. GLM `no_evolve` validation moves 45→40, 60→45, and 55→30. A later held-out move on that same version is a rescore, as in Table 27.

A.5 Per-round held-out record

Table 16 is every held-out pass in the 20-cell matrix. Regression is against the previous pass; round 0 is 0. Tokens are `total_tokens` summed on that pass’s held-out traces, not the 500,000 evolution-token cap. Validation is the stored `val_success`. Main-text regression rates are these values rounded to two decimals.

Table 16: Office-v1 held-out record. Success and validation are percent. Regression is a rate.

Model	Method	Seed	Round	Succ.	Val.	Reg.	Tokens
DS	ace	42	0	32.5	—	0.000	592020
DS	ace	42	1	30.0	40.0	0.231	690436
DS	ace	42	2	37.5	50.0	0.083	667912
DS	ace	42	3	37.5	35.0	0.267	808663

Model	Method	Seed	Round	Succ.	Val.	Reg.	Tokens
DS	ace	43	0	32.5	—	0.000	576054
DS	ace	43	1	25.0	30.0	0.308	634816
DS	ace	43	2	32.5	25.0	0.100	585464
DS	ace	43	3	47.5	35.0	0.077	958826
DS	ace	44	0	30.0	—	0.000	579534
DS	ace	44	1	27.5	30.0	0.167	594075
DS	ace	44	2	32.5	35.0	0.364	679860
DS	ace	44	3	32.5	35.0	0.385	967211
DS	expel	42	0	25.0	—	0.000	576918
DS	expel	42	1	32.5	20.0	0.200	704342
DS	expel	42	2	32.5	35.0	0.308	786634
DS	expel	42	3	42.5	35.0	0.154	776706
DS	expel	43	0	32.5	—	0.000	562761
DS	expel	43	1	30.0	35.0	0.077	671670
DS	expel	43	2	20.0	20.0	0.417	773513
DS	expel	43	3	42.5	25.0	0.000	673400
DS	expel	44	0	32.5	—	0.000	568396
DS	expel	44	1	27.5	25.0	0.231	680503
DS	expel	44	2	27.5	15.0	0.364	718343
DS	expel	44	3	50.0	30.0	0.091	694999
DS	gepa	42	0	35.0	—	0.000	580628
DS	gepa	42	1	25.0	25.0	0.357	590063
DS	gepa	42	2	40.0	30.0	0.200	685235
DS	gepa	42	3	42.5	35.0	0.250	635933
DS	gepa	43	0	30.0	—	0.000	571489
DS	gepa	43	1	30.0	30.0	0.167	555353
DS	gepa	43	2	40.0	35.0	0.167	692592
DS	gepa	43	3	30.0	20.0	0.375	736969
DS	gepa	44	0	25.0	—	0.000	608887
DS	gepa	44	1	35.0	30.0	0.100	556708
DS	gepa	44	2	35.0	20.0	0.071	578150
DS	gepa	44	3	42.5	25.0	0.143	609770
DS	no_evolve	42	0	35.0	—	0.000	563887
DS	no_evolve	42	1	32.5	20.0	0.286	577280
DS	no_evolve	42	2	27.5	20.0	0.308	619302
DS	no_evolve	42	3	45.0	25.0	0.091	511676
DS	no_evolve	43	0	27.5	—	0.000	611006
DS	no_evolve	43	1	35.0	30.0	0.182	575530
DS	no_evolve	43	2	32.5	30.0	0.143	598822
DS	no_evolve	43	3	35.0	20.0	0.231	564660
DS	no_evolve	44	0	25.0	—	0.000	612008
DS	no_evolve	44	1	30.0	30.0	0.200	597854
DS	no_evolve	44	2	22.5	25.0	0.333	616784
DS	no_evolve	44	3	40.0	30.0	0.000	534010
GLM	ace	42	0	47.5	—	0.000	457328
GLM	ace	42	1	47.5	50.0	0.105	431379

Model	Method	Seed	Round	Succ.	Val.	Reg.	Tokens
GLM	ace	42	2	35.0	45.0	0.316	695475
GLM	ace	42	3	37.5	—	0.071	689266
GLM	expel	42	0	37.5	—	0.000	466603
GLM	expel	42	1	50.0	50.0	0.133	415846
GLM	expel	42	2	35.0	60.0	0.400	677535
GLM	expel	42	3	37.5	—	0.071	698951
GLM	gepa	42	0	47.5	—	0.000	433726
GLM	gepa	42	1	45.0	50.0	0.211	582084
GLM	gepa	42	2	37.5	50.0	0.444	1142477
GLM	gepa	42	3	32.5	—	0.200	1229120
GLM	no_evolve	42	0	42.5	—	0.000	443087
GLM	no_evolve	42	1	35.0	45.0	0.235	464064
GLM	no_evolve	42	2	52.5	60.0	0.071	441988
GLM	no_evolve	42	3	35.0	55.0	0.381	659535
MM	ace	42	0	32.5	—	0.000	510135
MM	ace	42	1	20.0	15.0	0.538	775021
MM	ace	42	2	27.5	—	0.000	765450
MM	expel	42	0	22.5	—	0.000	523694
MM	expel	42	1	22.5	10.0	0.333	540754
MM	expel	42	2	20.0	15.0	0.222	533679
MM	expel	42	3	32.5	5.0	0.125	770089
MM	gepa	42	0	17.5	—	0.000	524370
MM	gepa	42	1	20.0	5.0	0.286	565039
MM	gepa	42	2	27.5	5.0	0.375	537457
MM	gepa	42	3	22.5	10.0	0.273	533844
MM	no_evolve	42	0	20.0	—	0.000	527353
MM	no_evolve	42	1	25.0	10.0	0.250	523358
MM	no_evolve	42	2	20.0	10.0	0.300	525339
MM	no_evolve	42	3	22.5	5.0	0.250	533712

A.6 DeepSeek family counts on every pass

Counts are solved held-out tasks. Column order is `catalog_discover` ($n=8$), `brief_report` (5), `checklist` (5), `csv_audit` (4), `access_denied` (6), `do_not_publish` (6). `incident_track`, `postmortem`, and `onboarding` are omitted: they are 0 on every DeepSeek pass.

A.7 Seed-42 family counts outside DeepSeek

DeepSeek family counts are Table 30 and Appendix A.6. The cells below are the seed-42 transfer. Entries are solved counts at round 0 and at the last recorded round.

Table 21: GLM 5.3 and MiniMax M2.5 solved counts, round 0 to last round.

Cell	Family	Round 0	Last
GLM no_evolve	access_denied	2/6	3/6
GLM no_evolve	brief_report	3/5	1/5
GLM no_evolve	catalog_discover	4/8	5/8
GLM no_evolve	checklist	2/5	0/5
GLM no_evolve	csv_audit	1/4	1/4
GLM no_evolve	do_not_publish	5/6	4/6
GLM gepa	access_denied	2/6	2/6
GLM gepa	brief_report	3/5	2/5
GLM gepa	catalog_discover	6/8	2/8
GLM gepa	checklist	3/5	2/5
GLM gepa	csv_audit	1/4	1/4
GLM gepa	do_not_publish	4/6	4/6
GLM ace	access_denied	2/6	1/6
GLM ace	brief_report	3/5	2/5
GLM ace	catalog_discover	6/8	6/8
GLM ace	checklist	2/5	1/5
GLM ace	csv_audit	1/4	1/4
GLM ace	do_not_publish	5/6	4/6
GLM expel	access_denied	1/6	2/6
GLM expel	brief_report	3/5	2/5
GLM expel	catalog_discover	3/8	6/8
GLM expel	checklist	1/5	0/5
GLM expel	csv_audit	2/4	1/4
GLM expel	do_not_publish	5/6	4/6
MM no_evolve	access_denied	1/6	1/6
MM no_evolve	brief_report	0/5	1/5
MM no_evolve	catalog_discover	5/8	4/8
MM no_evolve	checklist	1/5	0/5
MM no_evolve	do_not_publish	1/6	3/6
MM gepa	access_denied	1/6	1/6
MM gepa	brief_report	1/5	1/5
MM gepa	catalog_discover	2/8	3/8
MM gepa	checklist	2/5	1/5
MM gepa	do_not_publish	1/6	3/6
MM ace	access_denied	1/6	1/6
MM ace	brief_report	1/5	1/5
MM ace	catalog_discover	5/8	4/8
MM ace	checklist	2/5	1/5
MM ace	csv_audit	0/4	1/4
MM ace	do_not_publish	4/6	3/6
MM expel	access_denied	1/6	3/6
MM expel	brief_report	1/5	1/5
MM expel	catalog_discover	4/8	4/8

Cell	Family	Round 0	Last
MM expel	checklist	1/5	2/5
MM expel	csv_audit	0/4	1/4
MM expel	do_not_publish	2/6	2/6

Rows that stay at zero solved tasks on both ends are omitted. `incident_track`, `postmortem`, and `onboarding` are zero on every round of every seed-42 cell, matching the DeepSeek passes.

A.8 Failure reasons outside the DeepSeek office cells

Table 31 is DeepSeek office-v1. Table 22 is the seed-42 transfer on the same pack. Counts are `max_steps / verifier_fail`. A dash is a round the runner did not start.

Table 22: Seed-42 office-v1 transfer. Counts are `max_steps / verifier_fail`.

Cell	Seed	r0	r1	r2	r3
GLM no_evolve	42	17/6	22/4	14/5	22/4
GLM gepa	42	14/7	12/10	20/5	22/5
GLM ace	42	18/3	16/5	21/5	20/5
GLM expel	42	20/5	18/2	23/3	23/2
MM no_evolve	42	27/5	29/1	30/2	28/3
MM gepa	42	31/2	29/3	25/4	29/2
MM ace	42	23/4	29/3	29/0	—
MM expel	42	28/3	28/3	31/1	27/0

Every recorded failure in Table 22 is `max_steps` or `verifier_fail`.

Table 23: Manufacturing-v1 no_evolve. Counts are `below_threshold / max_steps`. These passes have no `verifier_fail`.

Seed	r0	r1	r2	r3
42	17/9	15/8	13/10	16/9
43	16/8	17/7	14/8	15/8
44	15/9	16/8	15/9	15/8

A.9 Eligibility flags

`gated` is false when final held-out regression exceeds 0.20, the run is over the evolution-token cap, the trace origin is not self, the anchor flag is not ok, the artifact does not match, contamination is set, M_0 is unbound, or the gate is off. `paper_min` repeats those gated reasons and also fails a pack that is under 40/10/20, has no OOD slice, overlaps split ids, or leaves the default mix and schedule. Coding has no OOD slice, so this runner reason fires there; the paper still reports that board and states the single family in Section 4. On these office-v1 cells the pack size is 80/20/40, so a `paper_min` mark that travels with `gated`

is the copied regression or budget reason, not a short pack. `main_table_eligible` can still be true for a frozen `no_evolve` cell whose regression is at most 0.20. That flag does not mean the policy changed. The Main column is the runner’s office-lane flag. It is not this paper’s ranking; coding is the primary scoreboard (Section 4).

Table 24: Stored eligibility reasons for the 20 office-v1 cells. Main is the runner’s office-lane flag, not the coding scoreboard.

Cell	Seed	Main	Reasons
DS ace	42	false	gated, paper_min
DS ace	43	true	eligible
DS ace	44	false	gated, paper_min
DS expel	42	true	eligible
DS expel	43	true	eligible
DS expel	44	true	eligible
DS gepa	42	false	gated, paper_min
DS gepa	43	false	gated, paper_min
DS gepa	44	true	eligible
DS no_evolve	42	true	eligible
DS no_evolve	43	false	gated, paper_min
DS no_evolve	44	true	eligible
GLM ace	42	false	gated, paper_min, over_budget
GLM expel	42	false	gated, paper_min, over_budget
GLM gepa	42	false	gated, paper_min, over_budget
GLM no_evolve	42	false	gated, paper_min
MM ace	42	false	gated, paper_min, over_budget
MM expel	42	true	eligible
MM gepa	42	false	gated, paper_min
MM no_evolve	42	false	gated, paper_min

A.10 Office grids

Setup. One hosted model is both the tool actor and the evolver. The actor is the TEB Standard Tool Actor: temperature 0, one user message, neutral prompt, at most 20 steps. Budget class is B0. The pack is frozen `office-v1`: catalog-workspace jobs, deterministic checks, no gold fields in the observation (all three packs: Appendix A.1). Evolution chunks are 27/27/26 of the 80 evolution tasks. Seeds {42, 43, 44} shuffle evolution order, not held-out membership. The three-seed methods were fixed from a development screen on validation gain versus model-matched `no_evolve`, then lower validation regression, then lower tokens, before formal held-out output was read: `gepa_llm`, `ace_llm` [Zhang et al., 2026b], `expel_llm`. Their mean screen gains were 0, −5.6, and −11.1 pp (Appendix A.3). `awm_llm` and `reasoningbank` [Ouyang et al., 2026] failed the screen’s artifact check. `dgm` and `clin_llm` [Majumder et al., 2024] ranked below and were not given this matrix. Every finished cell has `contamination=false` and `adapter_honesty=ok` (a labeled text update, not the upstream implementation).

What a pass is. The actor’s tools are `list_files`, `read_file`, `write_file`, `publish_artifact`, `grant_access`, and `submit`. A failed `submit` returns the failed check names and leaves the episode

open. Success is every terminal check passing. Gold files are not in the observation [Elangovan et al., 2021]. Four grade classes add checks on top of that contract: `discover` (names and list only), `extract` (bound phrases, length cap), `report` (facts under the required heading), and `structured` (strict JSON, no extra keys). Eight of the forty held-out tasks are `structured`. They stay inside the same Success denominator. Office Success is instruction following and rule-bounded delivery [Zhou et al., 2023, Ribeiro et al., 2020]. It is not a writing-quality score.

A.10.1 Frozen band

Table 25 and the first panel of Figure 7 are the control. DeepSeek v4 Pro and GLM 5.3 each span 17.5 pp with no context update. MiniMax M2.5 [MiniMax, 2026] spans 5.0 pp. DeepSeek seed 42’s path is 35.0, 32.5, 27.5, 45.0, all `agent_version=M0`. Actor, prompt, and held-out ids are identical across these passes; the hosted endpoint is not deterministic at temperature 0 [Atil et al., 2024], and seeds do not control it. Round 0 itself is 35.0, 27.5, and 25.0 across the three seeds, so Gain’s baseline is one draw from this band. Seed 43’s regression is 0.231, so the runner marks that frozen cell ineligible: the regression gate fires on rescore noise. On seed 42, every frozen MiniMax pass (20–25%) is below every frozen GLM pass (35–52.5%), so the pack separates models even where it cannot resolve updates.

A.10.2 Paired excess

Figure 7 shows that `no_evolve` moves with the adapters. Figure 8 shades the widest DeepSeek band, ± 17.5 pp, which is wider than seed 43’s own 7.5 pp band. Means and the per-seed reading are in Section 6. Table 11 is the cell record. Its Gain column still contains the frozen movement in Table 25.

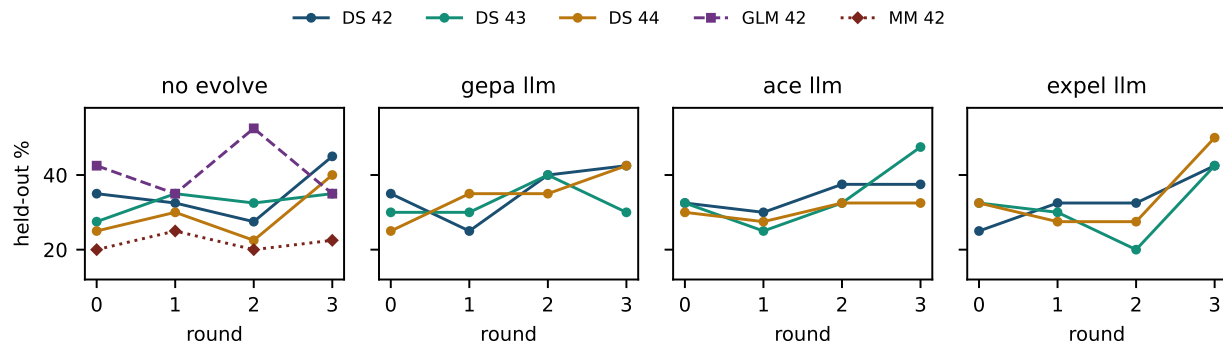


Figure 7: Held-out success by round. The first panel is frozen `no_evolve`, including GLM 5.3 and MiniMax M2.5 at seed 42. DeepSeek seed 42 goes from 27.5% at round 2 to 45% at round 3 with no context update. The other panels are DeepSeek only. `no_evolve` moves as much as the adapters.

A.10.3 Same agent, different score

A version change is an `evolve()` accept, including 0 pp ties under `min_delta=0`: DeepSeek `ace_llm` seed 42 (35→35 to M_3), `expel_llm` seed 44 (15→15, 30→30), and MiniMax `gepa_llm` (three ties to M_3 ; Appendix A.4). A score change without a version change is a rescore [Atil et al., 2024]. `expel_llm` seed 43 is M_2 at both round 2 (20.0%) and round 3 (42.5%): +22.5 pp on one artifact. `gepa_llm` seed 43 rescores M_1 from 40.0% to 30.0%. GLM `ace_llm` and `expel_llm` stay M_0 : rounds 1 and 2 roll back, and round 3 hits the evolution-token cap. Their held-out paths still span 12.5 pp and 15.0 pp. Table 27 lists

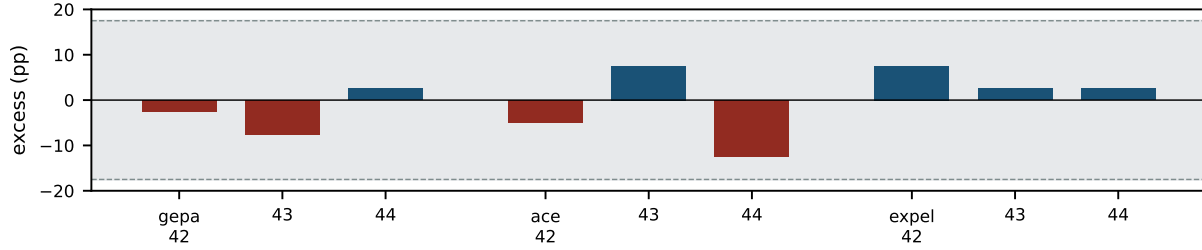


Figure 8: DeepSeek office paired excess. Shading is the widest seed band, ± 17.5 pp, not seed 43’s own 7.5 pp band. Each seed’s range is the Range column of Table 25.

same-version gaps of at least 10 pp. Table 26 is the accepted version after every round. Figure 9 is every finished cell, next to the full grid. MiniMax `ace_llm` has no round 3.

A.10.4 Full grid

Table 28 reports all 20 office-v1 cells; Figure 9 shows their round-wise scores.

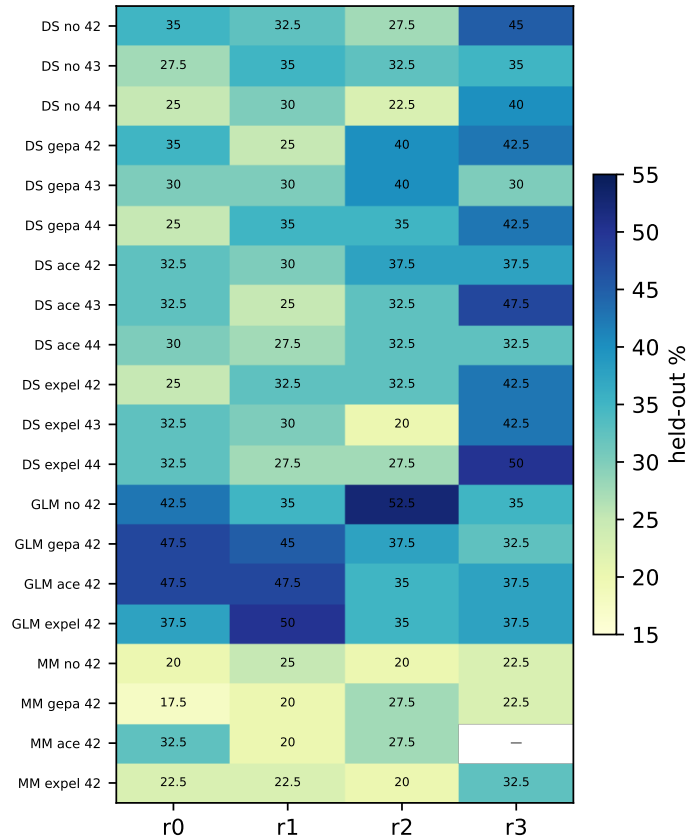


Figure 9: Held-out success (%) at each round for all 20 cells. A dash is a round the runner did not start (MiniMax `ace_llm` round 3).

Flags: reg. means final regression > 0.20 ; budget means `over_budget`. GLM rows that show a round-3 success still recorded that held-out pass and were then marked over the evolution-token cap. GLM `gepa_llm` regression is 0.200, which does not by itself exceed 0.20; the cell is ineligible because round 3 crossed 500,000 evolution tokens. DeepSeek `expel_llm` seeds 42 and 44 reach M_3 . Their excesses are +7.5 and +2.5 pp. Eligible DeepSeek excesses do not exceed the seed’s own range in Table 25. `ace_llm` seed 43 is +7.5 pp, equal to that seed’s 7.5 pp range. The other positive eligible excesses are `expel_llm` +7.5, +2.5, +2.5 and `gepa_llm` seed 44 +2.5, against the 17.5 pp ranges of seeds 42 and 44 (Appendix A.9).

A.10.5 Seed-42 transfer

MiniMax M2.5 and GLM 5.3 are one seed. The reading is in Section 6. Figure 10 and Table 29 are the cell record.

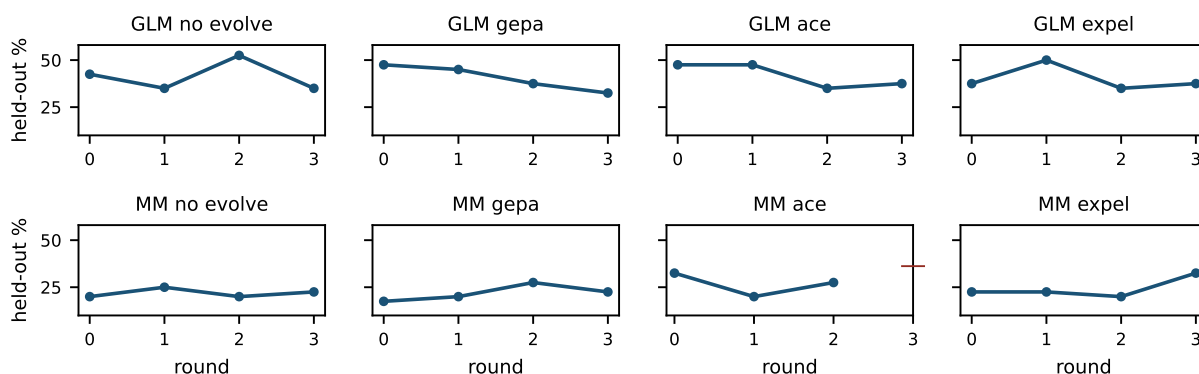


Figure 10: Seed-42 transfer. Top row: GLM 5.3. Bottom row: MiniMax M2.5. A dash is a round the runner did not start (MiniMax `ace_llm` round 3). These curves are not a second three-seed confirmation.

A.10.6 Where the 40 tasks move

Held-out $n=40$ splits into nine families (Table 30). Eight of the 80 evolution tasks are `content_draft`; this held-out fingerprint contains none, so Gain cannot show that family. `access_denied` and `do_not_publish` are held-out only ($n=6$ each) [Koh et al., 2021]. The other seven are not marked OOD on this held-out fingerprint. Figure 11 is the round-0 to final change in solved count for every DeepSeek cell. `incident_track` ($n=3$), `postmortem` ($n=2$), and `onboarding` ($n=1$) stay at zero solved tasks on every round of all twelve DeepSeek cells, including the three `expel_llm` runs that reach a new `agent_version`. The movement that produces raw Gain is among families the frozen policy already solves on some passes: `catalog_discover`, `brief_report`, `checklist`, `csv_audit`, `access_denied`, and `do_not_publish`.

On `no_evolve` seed 42 the final pass gains two `brief_report`, two `checklist`, one `catalog_discover`, and one `csv_audit`, and loses one `access_denied` and one `do_not_publish`. `expel_llm` seed 44, the highest final Success in the matrix (50%), moves `checklist` from 1/5 to 4/5 and `catalog_discover` from 5/8 to 7/8. Those are the same families that move when the agent string does not change. A family rate on $n \leq 8$ flips when one or two tasks flip. That is the 2.5 pp task grain inside the 17.5 pp band (Appendix A.17).

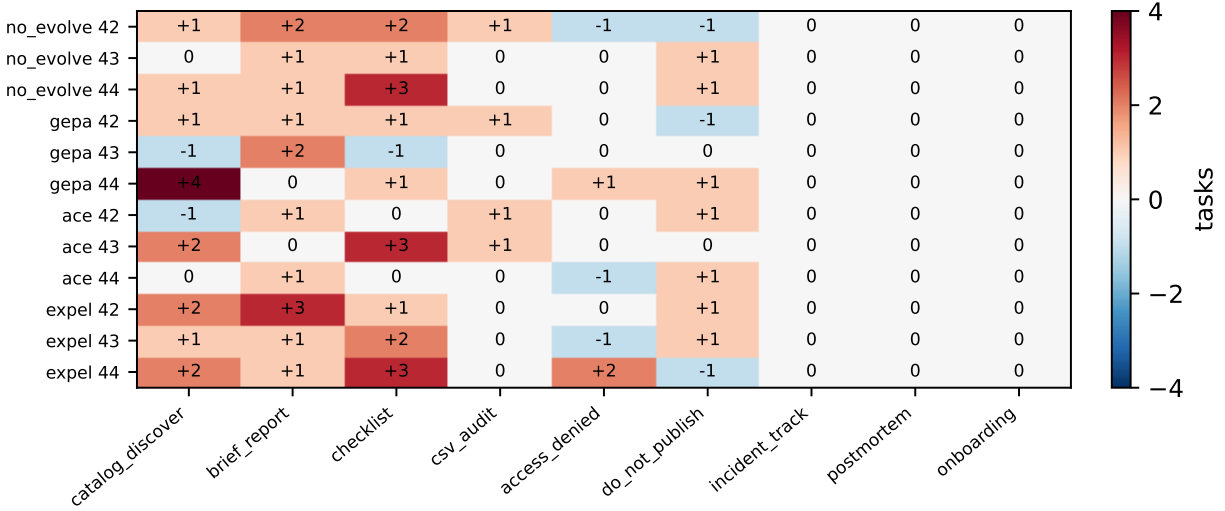


Figure 11: DeepSeek family change from round 0 to the final held-out pass, in solved tasks. Zero-change families `incident_track`, `postmortem`, and `onboarding` are included. Color is the task delta.

A.10.7 Failure reasons

Raw Gain ranks `no_evolve` seed 44 (+15 pp, M_0) with `gepa_llm` seed 44 and two `expel_llm` seeds at +17.5 pp. Paired excess ranks `gepa_llm` seed 43 and `ace_llm` seed 44 lowest (-7.5, -12.5). Table 31 is the movement: misses are the 20-step cap, and `no_evolve` seed 42 goes 23, 25, 25, 20 `max_steps` with no context update.

A.11 Office artifact versions

Table 28 in the main text is the score grid. Table 26 is the accepted artifact after each round. `no_evolve` stays at M_0 on every seed. A repeated version means the gate rolled back or the runner rescored the same artifact. GLM `ace_llm` and `expel_llm` stay at M_0 because the evolution-token cap skipped `evolve()`. GLM `gepa_llm` accepted M_1 and then rescored it. MiniMax `ace_llm` has no round 3.

The evolution-token cap is 500,000 per round. After each evolution task the runner adds that task’s token total, the sum of its steps. A step uses `usage.total_tokens` when the server sends it, otherwise prompt tokens plus completion tokens. Validation tasks, held-out tasks, and model calls inside `evolve()` are not added. Crossing the cap skips `evolve()` for that round, marks the cell `over_budget`, and skips later rounds when the violation is recorded before the next round.

Screen scores are not results. The screen ranked `gepa_llm`, `ace_llm`, and `expel_llm` ahead of `dgm` and `clin_llm` by mean validation gain against model-matched `no_evolve`, then lower validation regression, then lower token cost. `awm_llm` and `reasoning_bank` failed the screen’s artifact check. Those four methods were not given this held-out matrix. Running them on the same Gain definition would repeat the remeasurement in Table 25.

Every finished cell reports `contamination=false`, `adapter_honesty=ok`, `trace_origin=self`, `trace_mix=all`, `curriculum=none`, `budget_class=B0`, `human_edits=0`, and `m0_source=measured`. The held-out fingerprint is identical across the DeepSeek cells that share the frozen pack. Temperature is 0. The client sends one user message. Inference timeout is 60 seconds, with retries on transport errors and HTTP 429/502/503/504. HTTP 402 is not retried. Held-out failures on the DeepSeek `no_evolve` passes

are `max_steps` and verifier failures. A response cache of identical prompts would stabilize a rescore; it would not create the version changes in Table 26.

A.12 Manufacturing frozen repeat

Manufacturing-v1 uses the same 80/20/40 shape, $R=3$, and DeepSeek v4 Pro `no_evolve` on the Standard Tool Actor. The runner flags these cells `env_not_office`. This is the manufacturing frozen band, scored separately from coding. All three seeds stay at M_0 , `contamination=false`, `adapter_honesty=ok`, `over_budget=false`. The four-pass ranges are 7.5, 5.0, and 2.5 pp. Failures on these passes are `below_threshold` and `max_steps`; `verifier_fail` does not appear (Table 23). These ranges are not the office DeepSeek band and not the coding band (Appendix A.14). Office allows 20 steps per task; this repeat allows 12.

A.13 Manufacturing family counts

These are the finished DeepSeek v4 Pro `no_evolve` cells in Table 32. Each entry is solved tasks on that family at rounds 0/1/2/3. `agent_version` stays M_0 . The office lane rejects the cells (`env_not_office`). Families with zero solved tasks on every pass are `accounts_payable`, `advance_payments`, `advance_receipts`, `bpc_report`, `receivable_monthly`. `pc_revenue` is zero at both ends; seed 44 round 1 solves one task.

Table 33: Manufacturing-v1 solved counts by family at rounds 0–3, written r0/r1/r2/r3.

Family	n	Seed 42	Seed 43	Seed 44
<code>accounts_payable</code>	3	0/0/0/0	0/0/0/0	0/0/0/0
<code>advance_payments</code>	3	0/0/0/0	0/0/0/0	0/0/0/0
<code>advance_receipts</code>	1	0/0/0/0	0/0/0/0	0/0/0/0
<code>bpc_report</code>	7	0/0/0/0	0/0/0/0	0/0/0/0
<code>commodity</code>	3	1/2/2/2	2/2/2/2	2/2/2/2
<code>electricity</code>	3	3/3/3/3	3/3/3/3	3/3/3/3
<code>expense_detail</code>	4	1/2/1/1	1/1/2/2	2/1/1/2
<code>pc_revenue</code>	3	0/0/0/0	0/0/0/0	0/1/0/0
<code>purchase_receipt</code>	3	3/3/3/3	3/3/3/3	3/3/3/3
<code>receivable_monthly</code>	1	0/0/0/0	0/0/0/0	0/0/0/0
<code>revenue_cost</code>	3	1/1/2/0	1/1/2/2	0/1/1/1
<code>tax</code>	3	2/3/3/3	3/3/3/3	3/3/3/3
<code>utility</code>	3	3/3/3/3	3/3/3/2	3/2/3/3

A.14 Coding frozen repeat

DeepSeek v4 Pro `no_evolve` on the coding pack uses the same 80/20/40 shape and $R=3$. The runner flags these cells `env_not_office`. The four passes are Table 6. All three seeds stay at M_0 , `contamination=false`, `adapter_honesty=ok`, `over_budget=false`. The four-pass ranges are 7.5, 5.0, and 12.5 pp, with Gains of -7.5 , -2.5 , and -10.0 pp. The episode cap is 12 steps, not Office’s 20. Every miss is that cap.

Across the four passes the miss counts sum to 56, 60, and 63 on seeds 42, 43, and 44. Coding is a single family (`fix_bug`, $n=40$), so there is no family split to tabulate. Final regression is 0.231 and 0.308 on seeds 42 and 44: the 0.20 cutoff fires on frozen rescore noise here too, as on office seed 43.

A.15 Extra office frozen passes

Six extra DeepSeek `no_evolve` passes on the office-v1 held-out fingerprint (`afd748c6`, two per seed, $R=0$, no `evolve()` call) score 30.0 and 27.5 on seed 42, 27.5 and 22.5 on seed 43, and 30.0 and 27.5 on seed 44, all at M_0 with `contamination=false`. Merged with the original four passes, the per-seed ranges are 17.5, 12.5, and 17.5 pp. The widest stays 17.5 pp. Seed 43's own range widens from the 7.5 pp reported in Section 6 to 12.5 pp, because one extra pass scores 22.5. These are the extra repeats the Limitations ask for.

A.16 Coding grids

These displays use the same 36 coding cells as Section 4. Held-out is one family, so there is no family-delta figure. Every recorded miss is `max_steps`; `verifier_fail` does not appear.

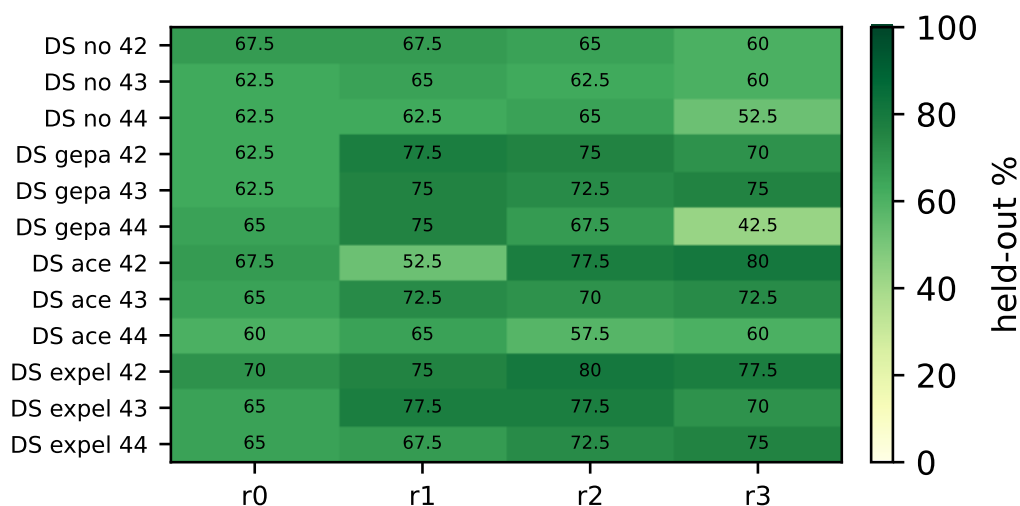


Figure 12: DeepSeek coding held-out success (%) at each round. Twelve cells: four methods by three seeds. Paired excess is Figure 4.

A.17 Worked examples

The three task cards are evolution-split tasks. They are not held-out items, and they omit gold files, expected SQL, and unit-test oracles. The last card is one stored DeepSeek `no_evolve` held-out trace pair; its instruction and written files are not reproduced.

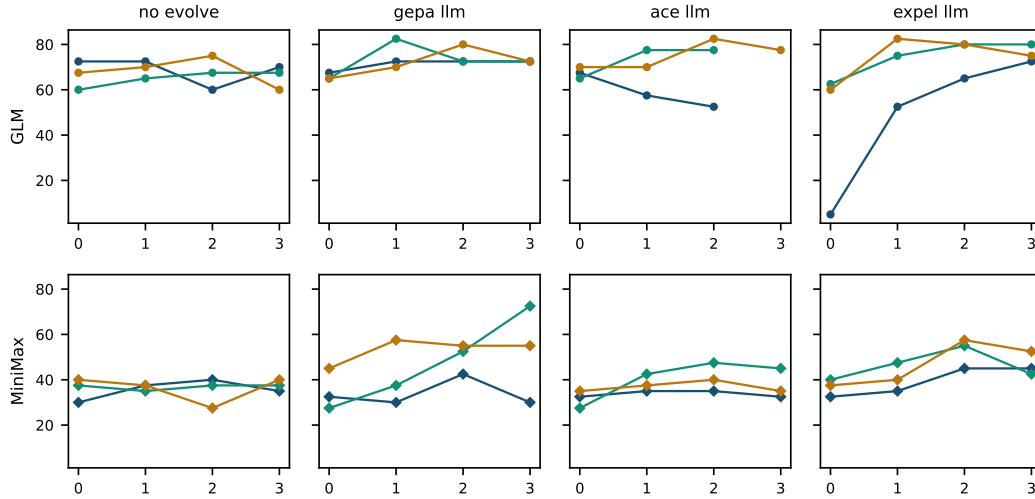


Figure 13: GLM (top row) and MiniMax (bottom row) coding held-out success by round. Both models have seeds 42, 43, and 44. Line color is the seed.

A.17.1 One task from each current suite

Office-v1 evolution task	
Task	<code>ofc.evo_discover_open_tasks</code>
Family	<code>catalog_discover</code> ; verification class <code>contract</code>
Instruction	“List iteration-12 files and publish <code>catalog_index.md</code> .”
Observation	Two catalog files: <code>catalog/iter12.md</code> , <code>catalog/owners.md</code> . The gold output/ <code>catalog_index.md</code> is not in the observation.
Tools	<code>list_files</code> , <code>read_file</code> , <code>write_file</code> , <code>publish_artifact</code> , <code>grant_access</code> , <code>submit</code>
Step cap	20
Success	Every terminal check passes. A failed <code>submit</code> returns the failed check names and leaves the episode open.

Coding-v1 evolution task	
Task	<code>coding_fix_hide_loading_gate</code>
Family	<code>fix_bug</code> ; verification class <code>executable</code>
Instruction	Chinese in the pack; English paraphrase: change <code>show_realname_gate</code> so a query that is still loading does not open the gate; only <code>required</code> or <code>unavailable</code> opens it.
Workspace	<code>src/solution.py</code>
Step cap	12
Success	Unit tests in the closed workspace pass. The gold patch is not shown.

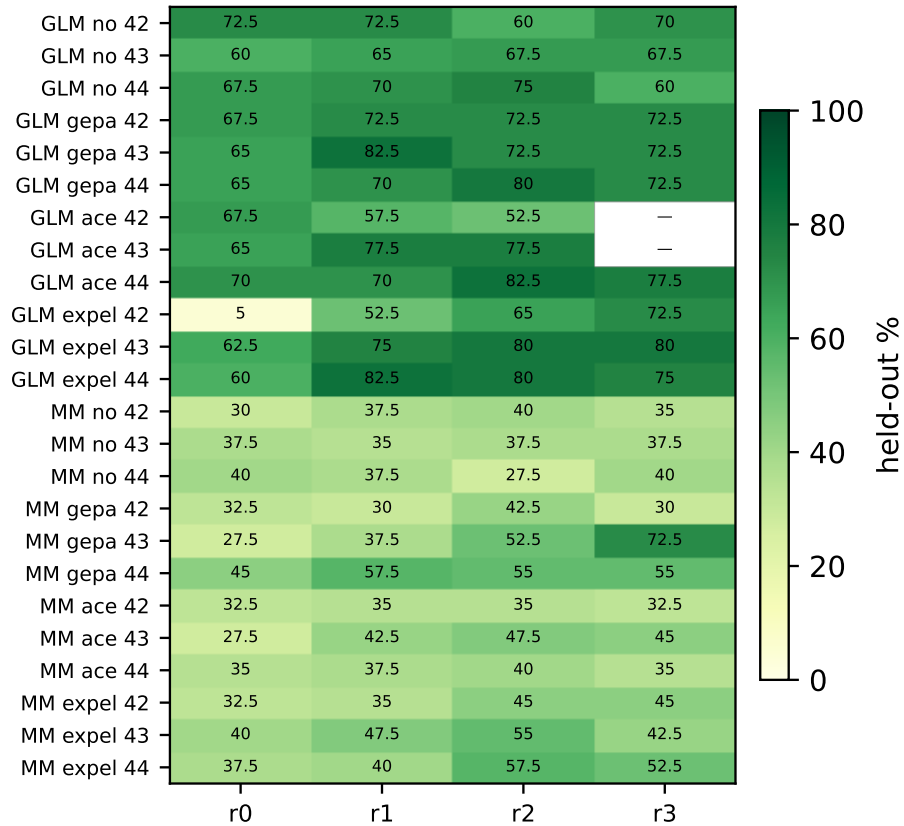


Figure 14: GLM and MiniMax coding held-out success (%) at each round. GLM `ace_llm` seeds 42 and 43 have no round 3. Paired excess for these cells is Figure 6.

Manufacturing-v1 evolution task	
Task	<code>manufacturing_x_tax_2025_personal</code>
Family	<code>tax</code>
Question	Chinese in the pack; English paraphrase: the group’s 2025 personal-income-tax total.
Data	Real-derived tables; counterparty, company, and department names pseudonymized.
Step cap	12 on this repeat
Success	Deterministic score at least 80. A miss is <code>below_threshold</code> or <code>max_steps</code> ; <code>verifier_fail</code> is not a manufacturing failure. The expected number and SQL are not shown.

A.17.2 What `evolve()` actually returns

Frozen control: DeepSeek <code>no_evolve</code> , office seed 42	
Update	None. <code>evolve()</code> returns the same agent.
Stored files	An evolve input for rounds 1–3; no evolve-artifact file.
Version	M_0 after every round.

0 pp accept: DeepSeek ace.11m, office seed 42, round 3

Gate	Validation 0.350 $\rightarrow$ 0.350 under min_delta=0; accept (Table 15).
Version	$M_2 \rightarrow M_3$
Carrier	Text playbook. Round 3 keeps the four round-2 bullets and adds a fifth.
Added bullet	Before the final <code>publish_artifact</code> that precedes submission, cross-check the draft against every task requirement and output rule, and edit the file to prevent <code>submit_constraints_failed</code> .

A.17.3 One held-out task, same M_0 , two scores

Rescore: DeepSeek no_evolve, office seed 42

Task	<code>ofc_hold_discover_rfq_archive (catalog_discover)</code>
Round 2	M_0 , 20 steps, fail (max_steps)
Round 3	M_0 , 20 steps, pass
Effect	One task is 2.5 pp. This pair is part of the seed-42 path from 27.5% at round 2 to 45% at round 3.

A.18 Archived ticket-gym mock

This section is not office-v1. The hosted matrix in Section A.10 uses catalog tools (`list_files`, `read_file`, `write_file`, `publish_artifact`, `grant_access`, `submit`) and held-out $n=40$. The numbers below are the earlier in-process ticket mock (held-out $n=20$). They are not ranked against Table 28.

Ticket-gym is a first-party in-process tool world: `search`, `read`, `write_file`, `update_ticket`, `auth_refresh`, `list_tickets`, `grant_access`, `submit`. The verifier is deterministic (files, ticket state, submit args), and the same `task_id` never crosses evolution / validation / held-out. Figure 15 visualizes the split; Table 38 lists its counts together with the inbox mock and the `office_work` expansion. Evolution and validation are in-distribution only; held-out adds six `access_denied` items (`ood=true`). Dual is therefore learnable from self failures; the OOD recover tool cannot be practiced on evolution.

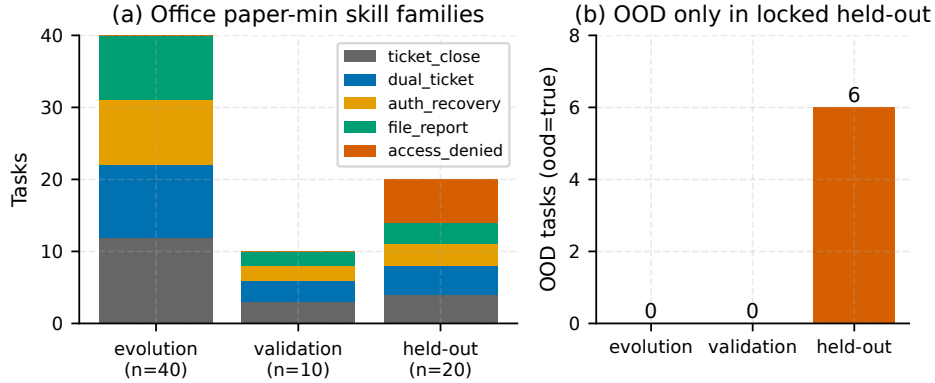


Figure 15: Archived ticket-gym, not office-v1. Split is 40/10/20 and held-out $n=20$. OOD is locked in held-out.

In-distribution families. `ticket_close`: find a policy, write a summary, close one ticket. `dual_ticket`: write a bundle covering two topics and close *two* tickets with distinct notes; an incomplete `submit` does

not end the episode. `auth_recovery`: `auth` can expire again after a successful refresh. `file_report`: report path plus required substrings.

OOD family. `access_denied` ($n=6$ on held-out only): a tool returns `permission_denied` until `grant_access`, which is *not* `auth_refresh`. Evolution never contains this family. Transfer here is “new recover tool,” not a new domain.

Canonical heuristic M_0 is ticket-close only, including `auth_refresh` on expiry. It does not know dual-ticket close or `grant_access`. On this pack that floor is 50% (10/20): the four ticket-close, three `auth`, and three file items, and none of dual or OOD.

A.18.1 Why the protocol constraints

Each check maps a provenance or evaluation failure to a stored eligibility flag.

A teacher dump on the same tasks can look like a large Gain with no successful self rollout. The NL2SQL paper-min pack is the control: teacher `schema_memory` reaches 100%, self traces stay at 0% (`anchor_floor`; Appendix A.18.14). Teacher and self traces therefore occupy separate labeled boards. If `evolve()` sees held-out ids or gold fields, Success is leakage: the runner hashes the sorted held-out `task_id` list and scans dumps. Training on held-out is not a TEB measurement.

Scoring a live server whose `artifact_id` differs from the value returned by `evolve()` breaks artifact binding. The discarded DGM-inspired run listed `grant_access_on_denied` in the mutation ISA and reached 80% / OOD 6/6 without OOD traces, so it is excluded (Appendix A.18.13). A B2 fullft and a B0 tip are different objects; the same evolver on a heuristic actor, frozen-3B wrapper, and native runtime is three cards.

Without rollback, wrapper GEPA [Agrawal et al., 2026] / Letta [Packer et al., 2023] / DGM [Zhang et al., 2026a] would look evolved; they reject every round, and `dgm_native` accepts with validation 70% while held-out stays 50% (Appendix A.18.12). Turning the gate off on B0 does not move those cards. $R=1$ cannot separate a jump from a loop; on this pack the heuristic jump is at `r1`, so $R=3$ is the eligibility floor. M_0 outside $[0.10, 0.80]$ cannot show Gain and is flagged.

The default mix is `all` (successes and failures) because failure traces carry information that success traces alone do not encode—on ticket-gym, dual-ticket learning relies on recoverable failed submits that expose the missing second ticket. Historical weight `success_only` artifacts are excluded with their parent card because their trace actor and evaluated actor differ; they cannot support a mix claim.

The default curriculum is `none` [Bengio et al., 2009] (static chunks) because static ordering keeps rounds comparable: every seed sees the same evolution ids in the same round, so round-level differences are attributable to the evolver, not to task selection. The `failure_family` ablation reorders remaining evolution ids toward families that failed on prior traces; it does not raise heuristic held-out because dual is already saturated and OOD never enters evolution. It is a labeled ablation, not a candidate default.

The default schedule is `accumulate_chunks`: each evolution id is collected in at most one round, and `evolve()` sees traces from rounds $1 \dots r$. `replay` reuses the first 14 evolution ids across all rounds instead; it is marked ineligible (`round_schedule`) because a second pass over the same tasks cannot introduce new failure modes. `accumulate_chunks` ensures that each round contributes new traces to the evolving payload, so a system that learns from novel failures is not starved of them.

The archived ticket-gym uses a 40/10/20 paper-min floor. Office-v1 and manufacturing-v1 are 80/20/40. That floor ($R=3$, ≥ 3 seeds) is a statistical minimum, not an optimum. Held-out $n=20$ gives each family at least 3–4 items so a family rate is not a single-task coin flip; evolution $n=40$ gives $R=3$ at least 13 new ids per round under `accumulate_chunks`, so a round is not a single-batch update. Fewer than these and a card’s Success % and Gain are noisy enough to obscure the family split that the protocol exists to expose.

TEB does not require weight updates or emit a scalar “evolution IQ.” Seeds do not alter held-out membership, failed self rollouts are not replaced with teacher traces, held-out inference is outside the update `budget_class`, and main-table verification is deterministic rather than LLM-as-judge [Zheng et al., 2023].

A.18.2 Success, Gain, and actors

Success % = $100 \times \#\{\text{held-out pass}\} / \#\{\text{held-out}\}$. Gain (pp) = $100 \times (\text{Succ}(M_r) - \text{Succ}(M_0))$, with M_0 *this-run* unless the card is eval-only with a frozen floor. Rank only inside (`env`, `budget_class`, `trace_origin`, `actor`). An adapter name is a retrofit; the actor is who emits the tool call.

A.18.3 Setup of the archived mock

All ticket-gym numbers use tasks 40/10/20, $R=3$, `trace_origin=self`, `mix all`, `schedule accumulate_chunks`, `gate min_delta=0`, unless labeled otherwise. Evolution collection is chunked at 14 tasks per round. Episode cap is 20 env steps. Seeds {42, 43, 44} shuffle evolution order, not held-out membership.

Heuristic B0. Actor: `SandboxAgent`. Evolvers: `no_evolve`, `prompt_memory`, `success_memory`, `stand_in_reflexion` [Shinn et al., 2023] / `expel` [Zhao et al., 2024] / `awm` [Wang et al., 2025]. No train GPU. Stand-in Reflexion maps failure traces onto executable tips (e.g. `workflow:dual_ticket_close`); there is no LLM judge.

Weight B2. Actor: Qwen2.5-3B-Instruct [Qwen Team, 2024] after self-trace full fine-tune. Eval binds `evolve`’s artifact. Gain vs. frozen heuristic $M_0=0.50$ (eval-only of a trained checkpoint is not a re-measure of untrained 3B M_0).

Verbal B0 wrappers. Same frozen 3B, TEB-stuffed tips, seed 42. Gain vs. *this-run* M_0 (measured 0.45/0.50/0.55 across cards; not the heuristic 0.50). Wrapper M_0 is unbound relative to the heuristic floor, so these rows are not ranked into the B0 heuristic card. These are not native runs of Packer et al. [2023], Agrawal et al. [2026], Zhang et al. [2026a], Zhai et al. [2025].

Native pairing (separate board): Letta [Packer et al., 2023] messages API as actor; DGM [Zhang et al., 2026a] genome compiled onto `SandboxAgent` without leaking OOD tool names in the mutation ISA; the same Reflexion-style [Shinn et al., 2023] wrapper loop on a larger chat model (`qwen-plus`).

`inbox` is a second first-party mock with the same 40/10/20 shape (Section A.18.8). The NL2SQL paper-min pack is a teacher-contrast negative control, not a third headline world (Appendix A.18.14).

A.18.4 Summary boards and expansion families

Table 39 separates the eligible heuristic ticket-gym card from one-seed wrapper contrasts and the incomplete weight attempt. Rows are boards, not methods to sort. The expansion pack is not another board: it has $M_0=0.20$ and different families, so 75% is not a denser 70%. Table 40 is that pack’s family counts. Table 48 is the same pack’s Success summary.

A.18.5 Round curves

Heuristic stand-in Reflexion jumps at r1 and saturates (70%). Other B0 methods stay at the M_0 floor. Among one-seed wrappers, only `reflexion_llm` moves (45→55); GEPA/Letta/DGM wrappers roll back to *this-run* M_0 ; `agent_evolver` 45→55 occurs under rollback and is M_0 jitter, not a second +10pp. Wrapper `file_report` stays 0/3.

A.18.6 Mechanism: what heuristic failures teach

At heuristic M_0 , ticket-close, auth, and file are already solved. Stand-in Reflexion’s +20pp is dual-ticket going 0/4→4/4 at r1 from *self* incomplete-submit traces; OOD stays 0/6 because that family is absent from evolution.

Figure 16 is a live SandboxAgent rollout on two locked held-out ids, not a 3B completion log. On dual (ofc_hold_du_001), M_0 searches once then repeatedly submits a single ticket; the verifier returns submit_constraints_failed and the episode continues to max_steps. After the stand-in writes workflow:dual_ticket_close, the same id is solved by two update_tickets with distinct notes. On OOD (ofc_hold_ad_001), M_0 never calls grant_access and dies in permission_denied; the recover tool is a one-step unlock, after which the ticket-close loop succeeds.

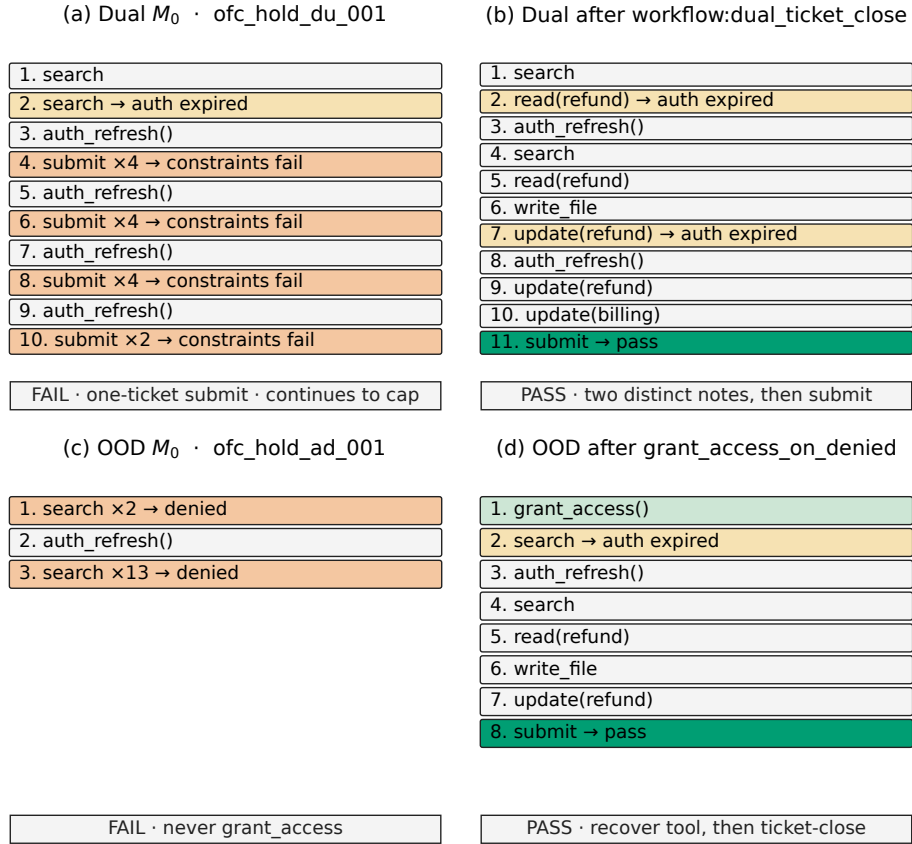


Figure 16: Archived ticket-gym tool calls, not office-v1. Locked held-out ids (SandboxAgent, not a 3B log). (a) M_0 dual: recoverable failed submit, second ticket never closed. (b) Same id after workflow:dual_ticket_close. (c) M_0 OOD: no grant_access. (d) OOD after grant_access_on_denied. Consecutive identical calls are collapsed ($\times n$).

A hygiene run that drops grant_access from the serving prompt puts no-hint 3B M_0 at 30%, OOD 0/6 (Table 44). The hinted wrapper’s M_0 OOD came from the prompt, not a learned recover skill. Frozen wrapper cards keep the hinted prompt so they reproduce; the no-hint M_0 is reported separately.

A.18.7 Why $R=3$ is enough on ticket-gym

Heuristic Reflexion jumps at r1 and saturates (70% / dual 4/4). Extra dual practice (`failure_family`) and `replay` of the first 14 evolution ids do not raise held-out, because dual is already solved for the stand-in and OOD `access_denied` is still absent from evolution. We therefore keep $R=3$ as the replication floor rather than grinding more rounds.

A.18.8 Inbox mock

To check that the protocol is not specific to ticket-gym, `inbox` provides a second in-process mock (`list`, `read`, `label`, `archive`, `reply`, `quarantine`) with the same 40/10/20 paper-min shape, $R=3$, and three seeds (Table 43). Canonical M_0 is triage-only (40%). B0 stand-in Reflexion reaches 70% / +30pp by unlocking `reply` 3/3 and thread 3/3, and never OOD `quarantine` (0/6). Other B0 methods stay at 40%.

As in ticket-gym, the stand-in learns an in-distribution family exposed by failed submits. Historical inbox weight grids are excluded because first-round traces came from `InboxAgent`, not the evaluated 3B. Inbox and ticket-gym remain separate boards; inbox is not a production mail client.

A.18.9 Seed, wrapper, inbox, and hygiene tables

These tables record the archived heuristic seeds, the seed-42 3B wrappers, the inbox board, and the prompt-hygiene check. They are not a ranking and not extra models. Step-token columns in the runner are `estimates` (Limitations); we do not treat them as a serving benchmark. Wall-clock is process time of the seed-42 3B loops only. Wrapper round paths below are held-out Success % of the *accepted* agent (rollback keeps M_{r-1} ; re-eval jitter is labeled). Native pairing is Table 46.

A.18.10 Weight-track provenance

Historical ticket-gym, inbox, and office-work weight grids are excluded. Their first round collected “self” traces with local heuristic actors, then trained and evaluated a 3B checkpoint; a scalar frozen heuristic M_0 did not bind that change of actor. Artifact matching alone cannot repair the provenance mismatch. The ticket-gym native attempt collects traces with the served 3B actor, but its `prepare_submit` round did not submit training and was rolled back; the resulting 50% / +0 card is an ineligible infrastructure outcome, not a weight-method result. We therefore report no weight-track headline.

A.18.11 Ticket-gym B0 ablations

Gate-off, replay, and failure-family runs keep heuristic Reflexion at 70% and other B0 methods at 50% (three seeds). Replay and failure-family remain labeled contrasts. Historical weight-mix and hybrid artifacts inherit the trace-actor provenance failure and are not ablation results.

A.18.12 Ticket-gym gate decisions

Figure 17 is the accept/rollback heatmap. Table 45 adds the held-out consequence. Green in the figure is accept; orange is rollback. Office seed 42 unless labeled B0. Validation Success is the gate input; held-out Success is scored on the accepted agent after that decision.

A.18.13 Ticket-gym native pairing

Not ranked vs. Table 39 (Table 46). `letta_native`: this-run M_0 60%, final 65% / +5pp, which is M2 re-eval jitter after r3 rollback (r1 held-out was 10%); dual 2/4 is the first 3B-backed dual on this pack, while

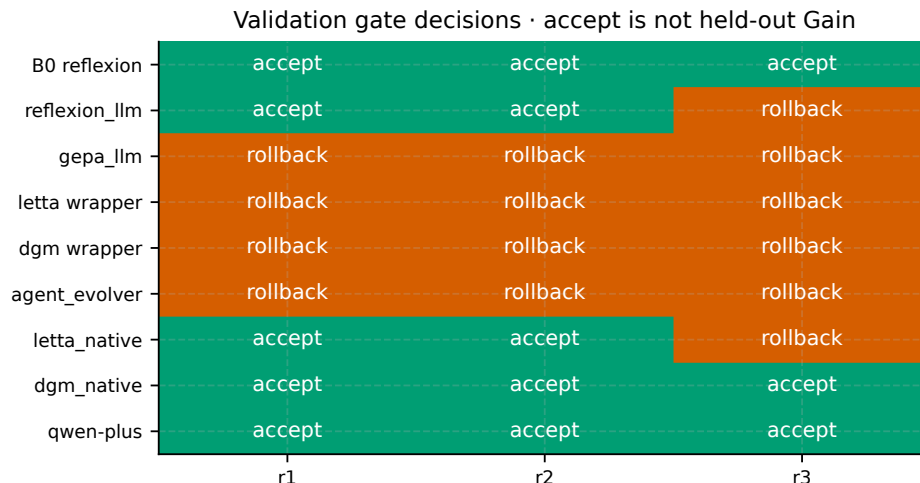


Figure 17: Archived ticket-gym gates, not office-v1. Accept and rollback are not held-out Gain.

ticket_close stays 0/4. dgm_native (ISA does not name OOD tools): 50% / +0pp. A discarded run that listed grant_access_on_denied in the mutation ISA hit 80% / OOD 6/6 without OOD traces—wrong mechanism, not a result. Historical weight grids are audit artifacts, not results.

The scale probe in this freeze is the same Reflexion-style wrapper loop on qwen-plus (a larger chat model than the bound 3B; not a 7B/14B fullft). This-run M_0 is 5%; peak M2 is 15%; the card is final 5% / +0pp / dual 0/4 (r3 undid M2). A bigger decoder did not unlock dual the way the heuristic stand-in does. We do not headline the M2 peak.

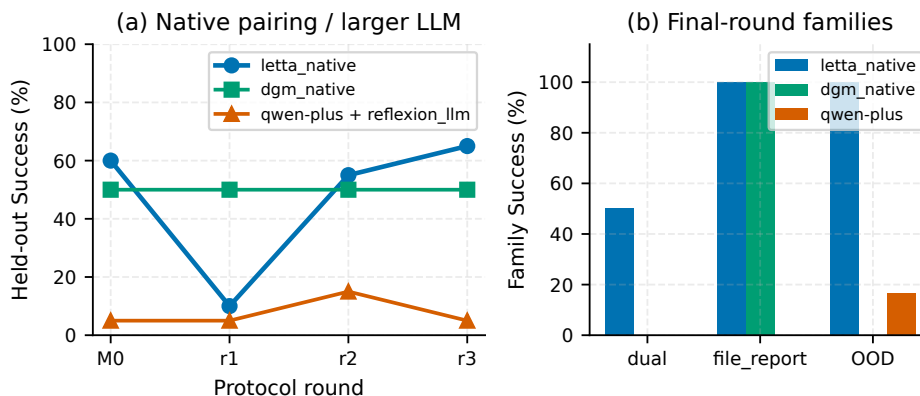


Figure 18: Archived ticket-gym native pairing board, not office-v1. Not ranked vs. Table 39.

A.18.14 NL2SQL teacher contrast

NL2SQL paper-min 40/10/20 [Yu et al., 2018, Li et al., 2023], not the 12/3/9 snapshot. Table 47: teacher schema_memory 0→100% by r1 (three seeds), OOD 4/4; self all 0% (anchor_floor). This is a labeled teacher contrast and a protocol floor, not an office-v1 result. Teacher 100% is not mixed with any self board.

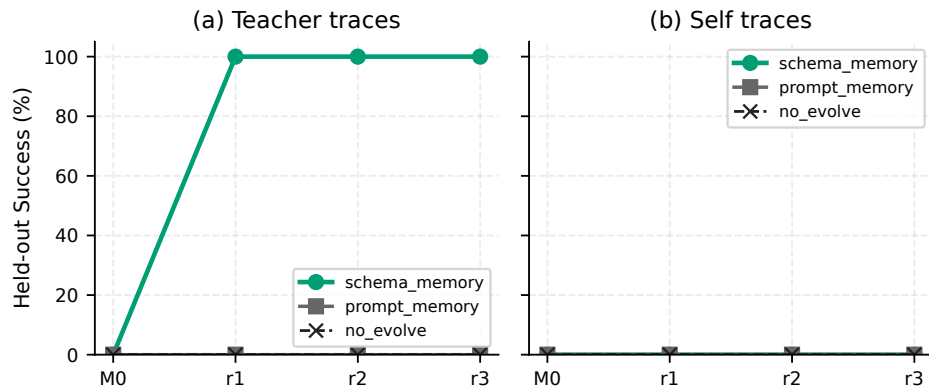


Figure 19: NL2SQL teacher contrast, not office-v1. (a) Teacher traces. (b) Self traces, which contain no successful SQL, so every evolver stays at 0%.

A.18.15 Ticket-gym expansion pack

Same ticket-gym tools, new families mapped from public product-usage cases, 40/10/20. Family counts are Table 40. Env id `office_work`, lane `contrast` (`pack_not_v1`). Not a fourth frozen actor board and not mixed with ticket-gym or inbox. Heuristic B0 (three seeds): measured M_0 20%; stand-in Reflexion 75% / +55pp (in-dist 15/15); other B0 stay 20%; OOD `access_denied` 0/3 and `checklist_bound` 0/2 (Figure 20, Table 48). No valid weight result is reported.

Table 15: Validation gate on office-v1. Columns are validation percent before and after the proposal. $n=20$, so one task is 5 pp.

Model	Method	Seed	Round	Version	Before	After	Decision
DS	ace	42	1	$M1$	20.0	40.0	accept
DS	ace	42	2	$M2$	35.0	50.0	accept
DS	ace	42	3	$M3$	35.0	35.0	accept
DS	ace	43	1	$M0$	30.0	20.0	rollback
DS	ace	43	2	$M0$	25.0	5.0	rollback
DS	ace	43	3	$M1$	15.0	35.0	accept
DS	ace	44	1	$M1$	25.0	30.0	accept
DS	ace	44	2	$M2$	25.0	35.0	accept
DS	ace	44	3	$M2$	35.0	25.0	rollback
DS	expel	42	1	$M1$	15.0	20.0	accept
DS	expel	42	2	$M2$	20.0	35.0	accept
DS	expel	42	3	$M3$	25.0	35.0	accept
DS	expel	43	1	$M1$	25.0	35.0	accept
DS	expel	43	2	$M2$	20.0	20.0	accept
DS	expel	43	3	$M2$	25.0	20.0	rollback
DS	expel	44	1	$M1$	15.0	25.0	accept
DS	expel	44	2	$M2$	15.0	15.0	accept
DS	expel	44	3	$M3$	30.0	30.0	accept
DS	gepa	42	1	$M0$	25.0	15.0	rollback
DS	gepa	42	2	$M1$	20.0	30.0	accept
DS	gepa	42	3	$M1$	35.0	25.0	rollback
DS	gepa	43	1	$M0$	30.0	25.0	rollback
DS	gepa	43	2	$M1$	20.0	35.0	accept
DS	gepa	43	3	$M1$	20.0	5.0	rollback
DS	gepa	44	1	$M0$	30.0	25.0	rollback
DS	gepa	44	2	$M0$	20.0	10.0	rollback
DS	gepa	44	3	$M1$	20.0	25.0	accept
DS	no_evolve	42	1	$M0$	20.0	15.0	rollback
DS	no_evolve	42	2	$M0$	20.0	20.0	accept
DS	no_evolve	42	3	$M0$	25.0	25.0	accept
DS	no_evolve	43	1	$M0$	30.0	15.0	rollback
DS	no_evolve	43	2	$M0$	30.0	15.0	rollback
DS	no_evolve	43	3	$M0$	20.0	20.0	accept
DS	no_evolve	44	1	$M0$	30.0	15.0	rollback
DS	no_evolve	44	2	$M0$	20.0	25.0	accept
DS	no_evolve	44	3	$M0$	30.0	25.0	rollback
GLM	ace	42	1	$M0$	50.0	40.0	rollback
GLM	ace	42	2	$M0$	45.0	35.0	rollback
GLM	ace	42	3	$M0$	—	—	over budget
GLM	expel	42	1	$M0$	50.0	35.0	rollback
GLM	expel	42	2	$M0$	60.0	30.0	rollback
GLM	expel	42	3	$M0$	—	—	over budget
GLM	gepa	42	1	$M1$	50.0	50.0	accept
GLM	gepa	42	2	$M1$	50.0	45.0	rollback
GLM	gepa	42	3	$M1$	—	—	over budget
GLM	no_evolve	42	1	$M0$	45.0	40.0	rollback
GLM	no_evolve	42	2	$M0$	60.0	45.0	rollback
GLM	no_evolve	42	3	$M0$	55.0	30.0	rollback
MM	ace	42	1	$M1$	10.0	15.0	accept
MM	ace	42	2	$M1$	—	—	over budget
MM	expel	42	1	$M0$	10.0	5.0	rollback
MM	expel	42	2	$M0$	15.0	10.0	rollback
MM	expel	42	3	$M1$	0.0	5.0	accept
MM	gepa	42	1	$M1$	5.0	5.0	accept
MM	gepa	42	2	$M2$	5.0	5.0	accept
MM	gepa	42	3	$M3$	10.0	10.0	accept
MM	no_evolve	42	1	$M0$	10.0	5.0	rollback
MM	no_evolve	42	2	$M0$	10.0	10.0	accept
MM	no_evolve	42	3	$M0$	5.0	0.0	rollback

Table 17: DeepSeek no_evolve solved counts by family and round. incident_track, postmortem, and onboarding are omitted (0 on every pass).

Seed	Round	<i>catalog-discover</i>	<i>brief-report</i>	<i>checklist</i>	<i>csv-audit</i>	<i>access-denied</i>	<i>do-not-publish</i>
42	0	6	0	0	1	2	5
42	1	4	1	1	1	2	4
42	2	4	0	0	1	1	5
42	3	7	2	2	2	1	4
43	0	5	1	0	1	1	3
43	1	6	0	1	1	2	4
43	2	5	0	1	1	2	4
43	3	5	2	1	1	1	4
44	0	5	0	0	1	1	3
44	1	4	1	1	1	2	3
44	2	4	0	0	1	1	3
44	3	6	1	3	1	1	4

Table 18: DeepSeek gepa_llm solved counts by family and round. incident_track, postmortem, and onboarding are omitted (0 on every pass).

Seed	Round	<i>catalog-discover</i>	<i>brief-report</i>	<i>checklist</i>	<i>csv-audit</i>	<i>access-denied</i>	<i>do-not-publish</i>
42	0	5	1	1	1	1	5
42	1	5	0	1	1	1	2
42	2	5	1	1	1	2	6
42	3	6	2	2	2	1	4
43	0	5	0	1	1	1	4
43	1	5	1	1	1	1	3
43	2	5	2	1	2	2	4
43	3	4	2	0	1	1	4
44	0	3	2	0	1	1	3
44	1	5	2	0	1	2	4
44	2	5	2	0	1	1	5
44	3	7	2	1	1	2	4

Table 19: DeepSeek ace_llm solved counts by family and round. incident_track, postmortem, and onboarding are omitted (0 on every pass).

Seed	Round	<i>catalog-discover</i>	<i>brief-report</i>	<i>checklist</i>	<i>csv-audit</i>	<i>access-denied</i>	<i>do-not-publish</i>
42	0	5	0	2	1	1	4
42	1	5	1	0	1	1	4
42	2	6	2	0	2	2	3
42	3	4	1	2	2	1	5
43	0	5	2	0	1	1	4
43	1	4	0	1	1	1	3
43	2	5	2	0	1	1	4
43	3	7	2	3	2	1	4
44	0	5	1	0	1	2	3
44	1	6	0	0	1	1	3
44	2	3	1	2	1	1	5
44	3	5	2	0	1	1	4

Table 20: DeepSeek expel_llm solved counts by family and round. incident_track, postmortem, and onboarding are omitted (0 on every pass).

Seed	Round	<i>catalog-discover</i>	<i>brief-report</i>	<i>checklist</i>	<i>csv-audit</i>	<i>access-denied</i>	<i>do-not-publish</i>
42	0	5	0	0	1	1	3
42	1	5	2	0	1	1	4
42	2	4	1	1	1	1	5
42	3	7	3	1	1	1	4
43	0	5	1	0	1	2	4
43	1	5	1	0	0	2	4
43	2	3	0	0	1	1	3
43	3	6	2	2	1	1	5
44	0	5	1	1	1	1	4
44	1	5	1	0	1	1	3
44	2	4	1	1	1	1	3
44	3	7	2	4	1	3	3

Table 25: Frozen-policy band. Range is max minus min of the four held-out passes, in percentage points.

Model	Seed	r0	r1	r2	r3	Range
DeepSeek v4 Pro	42	35.0	32.5	27.5	45.0	17.5
DeepSeek v4 Pro	43	27.5	35.0	32.5	35.0	7.5
DeepSeek v4 Pro	44	25.0	30.0	22.5	40.0	17.5
GLM 5.3	42	42.5	35.0	52.5	35.0	17.5
MiniMax M2.5	42	20.0	25.0	20.0	22.5	5.0

Table 26: Accepted `agent_version` after each office-v1 round. A repeated version means the artifact did not change. On `no_evolve` that is every round.

Model	Method	Seed	r0	r1	r2	r3
DS	no_evolve	42	M_0	M_0	M_0	M_0
DS	no_evolve	43	M_0	M_0	M_0	M_0
DS	no_evolve	44	M_0	M_0	M_0	M_0
DS	gepa	42	M_0	M_0	M_1	M_1
DS	gepa	43	M_0	M_0	M_1	M_1
DS	gepa	44	M_0	M_0	M_0	M_1
DS	ace	42	M_0	M_1	M_2	M_3
DS	ace	43	M_0	M_0	M_0	M_1
DS	ace	44	M_0	M_1	M_2	M_2
DS	expel	42	M_0	M_1	M_2	M_3
DS	expel	43	M_0	M_1	M_2	M_2
DS	expel	44	M_0	M_1	M_2	M_3
GLM	no_evolve	42	M_0	M_0	M_0	M_0
GLM	gepa	42	M_0	M_1	M_1	M_1
GLM	ace	42	M_0	M_0	M_0	M_0
GLM	expel	42	M_0	M_0	M_0	M_0
MM	no_evolve	42	M_0	M_0	M_0	M_0
MM	gepa	42	M_0	M_1	M_2	M_3
MM	ace	42	M_0	M_1	M_1	—
MM	expel	42	M_0	M_0	M_0	M_1

Table 27: Consecutive held-out passes with the same `agent_version` and an absolute gap of at least 10 pp. The accepted artifact did not change between the two scores.

Cell	Version	From	To	Gap
DS no_evolve 42	M_0	27.5	45.0	17.5
DS no_evolve 44	M_0	22.5	40.0	17.5
DS gepa 42	M_0	35.0	25.0	10.0
DS gepa 43	M_1	40.0	30.0	10.0
DS gepa 44	M_0	25.0	35.0	10.0
DS expel 43	M_2	20.0	42.5	22.5
GLM no_evolve 42	M_0	35.0	52.5	17.5
GLM no_evolve 42	M_0	52.5	35.0	17.5
GLM ace 42	M_0	47.5	35.0	12.5
GLM expel 42	M_0	37.5	50.0	12.5
GLM expel 42	M_0	50.0	35.0	15.0

Table 28: All 20 office-v1 cells. Success columns are held-out percent. Gain and excess are percentage points. Excess is blank on no_evolve. DS / GLM / MM are DeepSeek v4 Pro, GLM 5.3, and MiniMax M2.5.

Model	Method	Seed	r0	r1	r2	r3	Gain	Excess	Flag
DS	no_evolve	42	35.0	32.5	27.5	45.0	+10.0	—	
DS	no_evolve	43	27.5	35.0	32.5	35.0	+7.5	—	reg.
DS	no_evolve	44	25.0	30.0	22.5	40.0	+15.0	—	
DS	gepa	42	35.0	25.0	40.0	42.5	+7.5	-2.5	reg.
DS	gepa	43	30.0	30.0	40.0	30.0	0.0	-7.5	reg.
DS	gepa	44	25.0	35.0	35.0	42.5	+17.5	+2.5	
DS	ace	42	32.5	30.0	37.5	37.5	+5.0	-5.0	reg.
DS	ace	43	32.5	25.0	32.5	47.5	+15.0	+7.5	
DS	ace	44	30.0	27.5	32.5	32.5	+2.5	-12.5	reg.
DS	expel	42	25.0	32.5	32.5	42.5	+17.5	+7.5	
DS	expel	43	32.5	30.0	20.0	42.5	+10.0	+2.5	
DS	expel	44	32.5	27.5	27.5	50.0	+17.5	+2.5	
GLM	no_evolve	42	42.5	35.0	52.5	35.0	-7.5	—	reg.
GLM	gepa	42	47.5	45.0	37.5	32.5	-15.0	-7.5	budget
GLM	ace	42	47.5	47.5	35.0	37.5	-10.0	-2.5	budget
GLM	expel	42	37.5	50.0	35.0	37.5	0.0	+7.5	budget
MM	no_evolve	42	20.0	25.0	20.0	22.5	+2.5	—	reg.
MM	gepa	42	17.5	20.0	27.5	22.5	+5.0	+2.5	reg.
MM	ace	42	32.5	20.0	27.5	—	-5.0	-7.5	budget
MM	expel	42	22.5	22.5	20.0	32.5	+10.0	+7.5	

Table 29: Seed-42 transfer. Excess is against that model’s no_evolve. Not ranked with DeepSeek.

Model	Method	Succ.	Gain	Excess	Note
MiniMax	no_evolve	22.5	+2.5	—	reg. 0.25
MiniMax	gepa	22.5	+5.0	+2.5	inside 5.0; reg. 0.273
MiniMax	ace	27.5	-5.0	-7.5	stopped at round 2
MiniMax	expel	32.5	+10.0	+7.5	2.5 pp past a 5.0 band
GLM	no_evolve	35.0	-7.5	—	reg. 0.381
GLM	gepa	32.5	-15.0	-7.5	budget; reg. 0.20
GLM	ace	37.5	-10.0	-2.5	budget; version stays M_0
GLM	expel	37.5	0.0	+7.5	budget; version stays M_0

Table 30: Office-v1 held-out families. n sums to 40. OOD families do not appear in evolution.

Family	n	OOD	What the check is
catalog_discover	8	0	List /catalog, write and publish catalog.index.md
brief_report	5	0	Bounded summary; invented owners, dates, or claims fail
checklist	5	0	Required sections only
csv_audit	4	0	Recompute stated counts or clean a table
incident_track	3	0	Timeline, risk, and triage; unknowns stay unknown
postmortem	2	0	Facts and actions split; blame fails
onboarding	1	0	Numbered first-run steps
access_denied	6	6	Tools return permission.denied until grant.access
do_not_publish	6	6	Write the draft and submit without publish.artifact

Table 31: DeepSeek failure counts by round, `max_steps` / `verifier_fail`. The two counts sum to the number of held-out failures. No other failure reason appears.

Method	Seed	r0	r1	r2	r3
no_evolve	42	23/3	25/2	25/4	20/2
no_evolve	43	24/5	21/5	24/3	25/1
no_evolve	44	23/7	26/2	25/6	21/3
gepa	42	24/2	26/4	22/2	22/1
gepa	43	27/1	26/2	19/5	26/2
gepa	44	25/5	22/4	24/2	18/5
ace	42	21/6	24/4	21/4	22/3
ace	43	24/3	25/5	25/2	20/1
ace	44	23/5	25/4	26/1	22/5
expel	42	27/3	23/4	24/3	22/1
expel	43	24/3	24/4	24/8	20/3
expel	44	23/4	25/4	23/6	14/6

Table 32: Manufacturing-v1, DeepSeek v4 Pro, `no_evolve` only. Success is held-out percent. Version is M_0 at every round.

Seed	r0	r1	r2	r3	Range
42	35.0	42.5	42.5	37.5	7.5
43	40.0	40.0	45.0	42.5	5.0
44	40.0	40.0	40.0	42.5	2.5

Table 34: Coding frozen-policy passes for GLM and MiniMax. Gain and Range for all three models are Table 7. DeepSeek passes are Table 6. Range is max minus min of the four held-out passes, in percentage points.

Model	Seed	r0	r1	r2	r3	Gain	Range
GLM	42	72.5	72.5	60.0	70.0	-2.5	12.5
GLM	43	60.0	65.0	67.5	67.5	+7.5	7.5
GLM	44	67.5	70.0	75.0	60.0	-7.5	15.0
MiniMax	42	30.0	37.5	40.0	35.0	+5.0	10.0
MiniMax	43	37.5	35.0	37.5	37.5	+0.0	2.5
MiniMax	44	40.0	37.5	27.5	40.0	+0.0	12.5

Table 35: Accepted `agent_version` after each coding round. A repeated version means the artifact did not change. On `no_evolve` that is every round.

Model	Method	Seed	r0	r1	r2	r3
DeepSeek	no_evolve	42	M_0	M_0	M_0	M_0
DeepSeek	no_evolve	43	M_0	M_0	M_0	M_0
DeepSeek	no_evolve	44	M_0	M_0	M_0	M_0
DeepSeek	gepa	42	M_0	M_1	M_1	M_1
DeepSeek	gepa	43	M_0	M_1	M_1	M_1
DeepSeek	gepa	44	M_0	M_1	M_2	M_2
DeepSeek	ace	42	M_0	M_0	M_1	M_1
DeepSeek	ace	43	M_0	M_0	M_1	M_1
DeepSeek	ace	44	M_0	M_0	M_0	M_0
DeepSeek	expel	42	M_0	M_1	M_1	M_1
DeepSeek	expel	43	M_0	M_1	M_1	M_1
DeepSeek	expel	44	M_0	M_0	M_1	M_1
GLM	no_evolve	42	M_0	M_0	M_0	M_0
GLM	no_evolve	43	M_0	M_0	M_0	M_0
GLM	no_evolve	44	M_0	M_0	M_0	M_0
GLM	gepa	42	M_0	M_1	M_1	M_2
GLM	gepa	43	M_0	M_1	M_2	M_2
GLM	gepa	44	M_0	M_1	M_2	M_3
GLM	ace	42	M_0	M_1	M_1	—
GLM	ace	43	M_0	M_1	M_1	—
GLM	ace	44	M_0	M_1	M_1	M_2
GLM	expel	42	M_0	M_1	M_2	M_3
GLM	expel	43	M_0	M_1	M_2	M_2
GLM	expel	44	M_0	M_0	M_1	M_2
MiniMax	no_evolve	42	M_0	M_0	M_0	M_0
MiniMax	no_evolve	43	M_0	M_0	M_0	M_0
MiniMax	no_evolve	44	M_0	M_0	M_0	M_0
MiniMax	gepa	42	M_0	M_0	M_0	M_0
MiniMax	gepa	43	M_0	M_1	M_2	M_3
MiniMax	gepa	44	M_0	M_1	M_1	M_1
MiniMax	ace	42	M_0	M_1	M_2	M_3
MiniMax	ace	43	M_0	M_1	M_2	M_3
MiniMax	ace	44	M_0	M_1	M_2	M_3
MiniMax	expel	42	M_0	M_1	M_1	M_2
MiniMax	expel	43	M_0	M_1	M_1	M_1
MiniMax	expel	44	M_0	M_1	M_2	M_3

Table 36: Consecutive coding held-out passes with the same `agent_version` and an absolute gap of at least 10 pp. The accepted artifact did not change between the two scores.

Model	Method	Seed	Version	Rounds	From	To	Gap
DeepSeek	no_evolve	44	M_0	2–3	65.0	52.5	-12.5
DeepSeek	gepa	44	M_2	2–3	67.5	42.5	-25.0
DeepSeek	ace	42	M_0	0–1	67.5	52.5	-15.0
GLM	no_evolve	42	M_0	1–2	72.5	60.0	-12.5
GLM	no_evolve	42	M_0	2–3	60.0	70.0	+10.0
GLM	no_evolve	44	M_0	2–3	75.0	60.0	-15.0
GLM	ace	44	M_1	1–2	70.0	82.5	+12.5
GLM	expel	44	M_0	0–1	60.0	82.5	+22.5
MiniMax	no_evolve	44	M_0	1–2	37.5	27.5	-10.0
MiniMax	no_evolve	44	M_0	2–3	27.5	40.0	+12.5
MiniMax	gepa	42	M_0	1–2	30.0	42.5	+12.5
MiniMax	gepa	42	M_0	2–3	42.5	30.0	-12.5
MiniMax	expel	42	M_1	1–2	35.0	45.0	+10.0
MiniMax	expel	43	M_1	2–3	55.0	42.5	-12.5

Table 37: DeepSeek coding failure counts by round. The count is `max_steps.verifier_fail` is absent on every pass.

Method	Seed	r0	r1	r2	r3
no_evolve	42	13	13	14	16
no_evolve	43	15	14	15	16
no_evolve	44	15	15	14	19
gepa	42	15	9	10	12
gepa	43	15	10	11	10
gepa	44	14	10	13	23
ace	42	13	19	9	8
ace	43	14	11	12	11
ace	44	16	14	17	16
expel	42	12	10	8	9
expel	43	14	9	9	12
expel	44	14	13	11	10

Table 38: Archived ticket-gym family counts, held-out $n=20$, not office-v1. OOD families are held-out only. The same table lists the inbox mock and the `office_work` expansion.

Env	Family	Evo	Val	Held-out
ticket-gym	ticket_close	12	3	4
ticket-gym	dual_ticket	10	3	4
ticket-gym	auth_recovery	9	2	3
ticket-gym	file_report	9	2	3
ticket-gym	access_denied (OOD)	0	0	6
inbox	triage	16	4	8
inbox	reply_needed	12	3	3
inbox	thread_pair	12	3	3
inbox	quarantine (OOD)	0	0	6
office_work	ticket_close	6	2	4
office_work	issue_triage	8	2	3
office_work	release_notes	7	2	2
office_work	postmortem	7	2	2
office_work	diff_review	6	1	2
office_work	table_brief	6	1	2
office_work	access_denied (OOD)	0	0	3
office_work	checklist_bound (OOD)	0	0	2

Table 39: Archived ticket-gym mock, held-out $n=20$. Not an office-v1 score. Heuristic rows are three seeds. Wrapper rows are one seed. The weight attempt submitted no training.

Board	Succ.	Gain	dual	file	OOD	Seeds
B0 reflexion (heuristic)	70%	+20 vs 0.50	4/4	3/3	0/6	3
B0 other (heuristic)	50%	+0	0/4	3/3	0/6	3
Weight attempt (ineligible)	50%	+0 vs 0.50	0/4	3/3	0/6	1
3B reflexion.llm	55%	+10 vs 0.45	0/4	0/3	4/6	1
3B other wrappers	45–55%	+0 this-run	0/4	0/3	3–4/6	1

Table 40: Archived ticket-gym expansion pack `office_work`, not `office-v1`. Held-out $n=20$, three identical seeds. Heuristic B0 only. Not ranked vs. Table 39.

Family	n	M_0 / other B0	reflexion
<code>ticket_close</code>	4	4/4	4/4
<code>issue_triage</code>	3	0/3	3/3
<code>release_notes</code>	2	0/2	2/2
<code>postmortem</code>	2	0/2	2/2
<code>diff_review</code>	2	0/2	2/2
<code>table_brief</code>	2	0/2	2/2
<code>access_denied (OOD)</code>	3	0/3	0/3
<code>checklist_bound (OOD)</code>	2	0/2	0/2
Held-out Success	20	20%	75% (+55pp)

Table 41: Archived ticket-gym heuristic B0, not `office-v1`. Held-out Success % by seed. Other B0 methods are `no_evolve`, `prompt_memory`, `success_memory`, `expel`, `awm`.

Evolver	s42	s43	s44	dual	OOD
<code>reflexion (stand-in)</code>	70	70	70	4/4	0/6
<code>no_evolve</code>	50	50	50	0/4	0/6
<code>prompt_memory</code>	50	50	50	0/4	0/6
<code>success_memory</code>	50	50	50	0/4	0/6
<code>expel</code>	50	50	50	0/4	0/6
<code>awm</code>	50	50	50	0/4	0/6

Table 42: Archived ticket-gym 3B tip-wrappers, not `office-v1`. Seed 42. Columns `r0`–`r3` are held-out Success % of the accepted agent. `reflexion_llm` `r3` rolls back (card is M2 55%). `agent_evolver` `45`→`50`→`55` occurs under three rollbacks (jitter), not a second +10pp. Dual 0/4 and file 0/3 on every row. Wall is loop process time, not time-to-first-token.

Loop	r0	r1	r2	r3	Gain	OOD	Wall (h)
<code>reflexion_llm</code>	45	55	55	55	+10	4/6	4.0
<code>gepa_llm</code>	50	50	50	50	+0	3/6	3.8
<code>letta wrapper</code>	55	55	55	55	+0	4/6	5.9
<code>dgm wrapper</code>	45	45	45	45	+0	3/6	5.7
<code>agent_evolver</code>	45	50	55	55	jitter	4/6	5.4

Table 43: Archived inbox mock, not `office-v1`. Held-out $n=20$. Not ranked vs. Table 39. Reply $n=3$, thread $n=3$, OOD quarantine $n=6$.

Board	Succ.	Gain	reply	thread	OOD	Seeds
B0 reflexion	70%	+30 vs 0.40	3/3	3/3	0/6	3
B0 other	40%	+0	0/3	0/3	0/6	3

Table 44: Archived ticket-gym hygiene, not office-v1. Same frozen 3B, serving prompt with vs. without `grant_access` hint. Eval-only ($R=0$). Dual 0/4 on both rows. Table 42 keeps the hinted prompt.

Prompt	Succ.	tc	auth	file	OOD
Hinted wrapper M_0	45%	4/4	2/3	0/3	3/6
No hint M_0	30%	4/4	2/3	0/3	0/6

Table 45: Archived ticket-gym gates, not office-v1 (Table 15). A = accept, R = rollback. `dgm_native` can accept a validation 70% proposal while held-out stays 50%. `qwen-plus` r3 accepts 0→0; the card is final 5%, not the M2 peak.

Card	r1	r2	r3	Held-out note
B0 reflexion	A	A	A	70% from r1
reflexion_llm	A	A	R	stays M2 55%
gepa/letta/dgm wrap	R	R	R	this-run M_0
agent_evolver	R	R	R	45→55 jitter
letta_native	A	A	R	r1 is 10%; final M2 jitter 65%
dgm_native	A	A	A	val 70%, held-out 50%
qwen-plus	A	A	A	r0-r3: 5, 5, 15, 5

Table 46: Archived ticket-gym native pairing, not office-v1. Seed 42. Columns r0–r3 are held-out Success % of the accepted agent. Dual/OOD are the frozen (final) card. Not ranked vs. Table 39. `letta_native` r3 rolls back: final 65% is M2 re-eval jitter vs. r2 55%. `qwen-plus` peak r2 15% is not the card. `dgm_native` actor is `SandboxAgent` (genome only). File: Letta/DGM 3/3, `qwen-plus` 0/3. Ticket-close: Letta 0/4, DGM 4/4, `qwen-plus` 0/4.

Card	r0	r1	r2	r3	Gain	dual	OOD
letta_native	60	10	55	65	+5 jitter	2/4	6/6
dgm_native	50	50	50	50	+0	0/4	0/6
qwen-plus	5	5	15	5	+0	0/4	1/6

Table 47: NL2SQL teacher contrast, not office-v1. Held-out Success %, mean of seeds 42–44 (cells identical across seeds). Self has no successful SQL in traces.

Evolver	Teacher traces	Self traces
schema_memory	100%	0%
prompt_memory	0%	0%
no_evolve	0%	0%

Table 48: Archived ticket-gym `office_work` heuristic B0, not office-v1. Held-out Success %. Identical across seeds 42–44. Other B0: `no_evolve`, `prompt_memory`, `success_memory`, `expel`, `awm`.

Evolver	Success	Gain	in-dist	OOD
reflexion (stand-in)	75%	+55pp	15/15	0/5
other B0	20%	+0pp	4/15 (tc only)	0/5

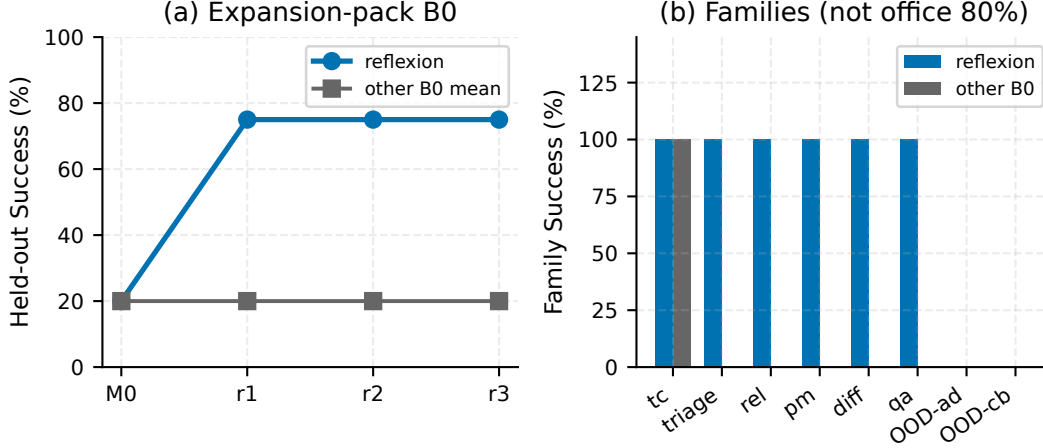


Figure 20: Archived ticket-gym expansion pack, not office-v1. Not ranked vs. Table 39.

A.18.16 Cross-environment summary

Table 49 consolidates the frozen ticket-gym, inbox, NL2SQL, and `office_work` cards. Rows are actor boards. Each row uses the 40/10/20 pack, $R=3$, the validation gate, and artifact binding. Seed counts are the last column: the 3B wrapper row is one seed, and the other rows are three. No single row is ranked above another across budget classes or environments.

Table 49: Archived cards across ticket-gym, inbox, NL2SQL, and `office_work`. Not an office-v1 ranking. “—” = family not present in that environment.

Env	Board	Budget	Succ.	Gain	In-dist family	OOD family	Seeds
ticket-gym	B0 reflexion (heur.)	B0	70%	+20 vs 0.50	dual 4/4	0/6	3
	3B reflexion.llm wrap	B0	55%	+10 vs 0.45	dual 0/4	4/6	1
inbox	B0 reflexion (heur.)	B0	70%	+30 vs 0.40	reply 3/3, thread 3/3	0/6	3
	B0 other (heur.)	B0	40%	+0	0/3	0/6	3
NL2SQL (contrast)	Teacher schema_mem	—	100%	+100 vs 0	OOD 4/4	—	3
	Self (all evolvers)	—	0%	—	0/4	—	3
office_work	B0 reflexion (heur.)	B0	75%	+55 vs 0.20	in-dist 15/15	0/5	3