Evaluating LLM Agents Beyond Local Edit Validity

Anonymous ACL submission

Abstract

Sequential large language model (LLM) evaluations often ask whether agents emit valid local actions or eventually succeed, but delayed constructive tasks also require *recoverability*: after each accepted edit, can the partial artifact still be completed within the remaining budget? We introduce a recoverability-centered measurement protocol over a shared edit interface spanning Word Ladder, alloy-like composition editing, and restricted protein G B1 (GB1) sequence-landscape editing. The environments form a control-to-mechanism ladder: Word Ladder exposes interface validity, Alloy-like editing tests delayed trade-off management, and GB1 permits exact auditing under deceptive local evidence. Across six models, Word Ladder has low local validity (36.1%), whereas Alloy-like editing and GB1 are often surface-clean but have low surface-clean success rates (16.0% and 6.8%). In GB1, exact enumeration over 625 variants and budgets 0–6 validates the recoverability probe: 51.4% of surface-clean episodes cross an exact boundary, and residual non-crossing success is 14.1%. Alloy-like audits give bounded evidence of recoverability drops and rounded-state boundary loss, while wrapper and MCTS checks show which interventions target each bottleneck. Recoverability gives sequential agent evaluation a decomposable, auditable, intervention-informative reporting axis.

1 Introduction

Large language models (LLMs) are increasingly evaluated as agents that act over multiple steps (Liu et al., 2024; Zhou et al., 2024; Wang et al., 2022b; Xie et al., 2024; Yao et al., 2023b). A common question is whether the model can emit a well-formed and admissible next action. For constructive tasks, this is necessary but not sufficient: a locally valid edit can consume the remaining budget, commit to an incompatible partial solution, or move into a local optimum. The central question is whether the resulting partial solution remains recoverable, meaning that some feasible completion still exists under the remaining budget.

We use **constructive structure editing** as the measurement substrate: an agent incrementally modifies a partial structured output through local edits, under local constraints, while final success is determined by delayed global evaluation. This connects to structured prediction as search, iterative editing, local search, and sequential decision making under delayed feedback. The contribution is not a new planning formalism; it is a recoverability-centered reporting method that separates local action admissibility from trajectory-state preservation.

A partial solution is recoverable if the current state and remaining edit budget admit a path to satisfying the delayed objective. We operationalize this target with environment-specific recoverability probes used only for analysis, not as model-facing privileged signals. For GB1, the restricted 625-variant landscape also permits exact enumeration, which validates the probe and audits stored trajectories. We call an episode **surface-clean** when every action is well formed and locally admissible. Surface-clean failure is therefore the key phenomenon for the composition claims: the model has satisfied the interface, but the trajectory still fails the delayed objective. Final success alone cannot distinguish interface failure, poor stopping, local deception, and recoverable-but-not-found search failure.

We instantiate one edit protocol in three controlled roles. Word Ladder is an action-interface control. Alloy-like composition editing shifts the bottleneck to delayed trade-offs between ultimate tensile strength (UTS), density, and recovery cost. GB1 landscape editing is the primary exact recoverability case: local additive evidence can be deceptive because final fitness depends on

Shared protocol	Observe partial structured output	Apply locally admissible edit	Delayed global evaluation
Word Ladder	word path in a Hamming graph	one-character lexical edit	target reachability under budget
Alloy-like	fixed-mass composition vector	mass transfer between elements	delayed strength–density feasibility
GB1 landscape	restricted GB1 sequence variant	one-position token edit	epistatic fitness threshold

Figure 1: **Constructive structure editing as a shared edit substrate.** Across environments, the model edits a partial structured output through local actions and receives delayed global evaluation. The environments differ in the source of trajectory difficulty: graph reachability, multi-objective trade-off, and deceptive sequence-landscape interactions.

epistatic interactions. Alloy-like editing and GB1 are controlled proxies for interaction structure under a matched local-edit interface, not materials-discovery or protein-engineering benchmarks.

Our main finding is hierarchical. Word Ladder shows why surface validity must be reported; Alloy-like editing and GB1 show the stronger composition failure: many episodes are surface-clean, yet surface-clean success remains only 16.0% and 6.8%, respectively. Exact GB1 enumeration decomposes this failure rate into boundary loss and residual failure despite retained recoverability. Prompt, history, wrapper, and Monte Carlo Tree Search (MCTS) checks then become targeted stress tests: recoverability measurements indicate whether format repair, stopping checks, visible-state search, backtracking, or privileged objective information is the plausible repair.

Contributions. We make four contributions. First, we formalize recoverability-centered reporting for constructive sequential agents and give an exact decomposition of surface-clean success where recoverability labels are available. Second, we instantiate a matched edit substrate spanning lexical, compositional, and restricted sequence-landscape editing, arranged as a control-to-mechanism ladder: interface failure in Word Ladder, bounded trajectory-management evidence in Alloy-like editing, and exact recoverability auditing in GB1. Third, we show that surface-clean, locally admissible trajectories often fail delayed objectives in Alloy-like editing and GB1, with GB1 failures decomposed by exact state-budget enumeration. Fourth, we use prompt, history, wrapper, and MCTS checks to show that the measured failure mechanism constrains the plausible intervention class, and release trajectory-level labels, recoverability diagnostics, exact GB1 validation, Alloy trajectory-state audits, raw logs, and verification scripts.

2 Related Work

LLM agents and sequential evaluation. Agent benchmarks evaluate LLMs in embodied, web, game, computer-use, software, ML-engineering, and scientific settings (Liu et al., 2024; Zhou et al., 2024; Wang et al., 2022b; Shridhar et al., 2021; Deng et al., 2023; Yao et al., 2022; Xie et al., 2024; Paglieri et al., 2025; Jimenez et al., 2024; Chan et al., 2025). These suites are valuable for end-to-end capability measurement, but their breadth can mix instruction following, tool knowledge, domain knowledge, local action validity, and longer-horizon trajectory management. Our protocol is intentionally narrower: it fixes the action surface and measures whether locally admissible edits preserve a feasible completion.

Structured prediction, editing, and search. The abstraction is closest to structured prediction as search over partial outputs (Doppa et al., 2014), iterative editing for text transformation (Guu et al., 2018; Malmi et al., 2019; Gu et al., 2019), and local search or combinatorial optimization where myopic improvements can create local optima (Kokel et al., 2025; Gonzalez-Pumariega et al., 2025; Sun et al., 2026). Inference-time methods such as self-consistency, tree search, ReAct, Reflexion, and Self-Refine aim to improve reasoning or refinement (Wang et al., 2023; Yao et al., 2023a,b; Shinn et al., 2023; Madaan et al., 2023). We do not claim to close the space of planning or search interventions. Instead, we provide a measurement setting where such interventions can be evaluated by whether they preserve feasible completion, not only by whether they emit admissible actions.

Controlled scientific-design-inspired proxies. Materials and biological sequence design involve local changes whose consequences depend on delayed properties, multi-objective trade-offs, or rugged fitness landscapes (Sánchez-Lengeling and Aspuru-Guzik, 2018; Wang et al., 2022a; Borg et al., 2020; Du et al., 2023; Dallago et al., 2021; Sandhu et al., 2025). We use alloy-like composition editing and GB1 landscape editing as controlled sources of such interaction structure. The claim is not that the environments measure deployable scientific-design capability. Their role is to stress feasible-completion preservation under a simple local-edit grammar.

Across these lines of work, our reframing is measurement-oriented rather than model-architectural: local action validity is necessary but not sufficient for sequential agents, and recoverability is the trajectory-level quantity we make observable and, in GB1, decomposable.

3 Method

Recoverability-centered constructive editing

For each environment e, define a task $\mathcal{T}_e = (\mathcal{S}_e, \mathcal{A}_e, P_e, O_e, G_e, H_e)$. At step t, the model observes $o_t = O_e(x_t, \tau_{<t})$, where x_t is the current partial structured output and $\tau_{<t}$ is the edit history. It either emits a local edit $a_t \in \mathcal{A}_e^{\text{edit}}$, which updates the state by $x_{t+1} = P_e(x_t, a_t)$, or emits `FINALIZE`, which triggers delayed global evaluation $G_e(\tau_{\leq t}, x_t)$. Success depends on the final state and, when relevant, trajectory-induced quantities: $S(\tau, x_T) \in \{0, 1\}$. Word Ladder also terminates successfully when an accepted edit reaches the target; we treat this as an environment-specific terminal evaluation. This is a finite-horizon sequential decision problem with partial observations: the agent sees text observations and local feedback, but not the full delayed objective or the exact future recoverability table. The constructive-editing terminology is therefore not a new planning formalism. It is the edit substrate on which we define the recoverability measurements below: the action-interface and trajectory-state parts of the decision problem are logged separately, so they can fail separately.

We distinguish four concepts. A **well-formed action** matches the parser. A **locally admissible edit** is accepted by the environment from the current state. A **surface-clean episode** contains no malformed or locally invalid actions. A **feasible completion** exists when the current partial solution still admits a successful completion under the remaining edit budget. **Recoverability** is the measured or estimated version of this last property. It is not a general planning metric and is never shown to the model. In GB1, the finite restricted landscape allows exact enumeration, which we use to validate this diagnostic for the main sequence-landscape claims.

This gives the decomposition we report. Let C denote a surface-clean episode, S terminal success, and B an exact pre-terminal recoverability-boundary crossing. When B is exact, crossing it precludes success, so

$$P(S \mid C) = P(\neg B \mid C)\, P(S \mid C, \neg B). \quad (1)$$

In the six-model GB1 audit, 51.4% of surface-clean episodes cross this exact boundary and residual non-crossing success is 14.1%, yielding the reported 6.8% surface-clean success. For alloy-like editing, B is not exact: recoverability drops remain bounded-probe degradation signals.

Environments

Word Ladder: graph-constrained lexical editing. The state contains the current word, target word, visited words, and remaining budget. An edit changes one character and is locally admissible only if the new word is in the lexicon, differs by one character, and has not been visited. The main bottleneck is the action interface plus reachability under budget: the environment supplies the control case where low local validity and budget slack are visible before stronger composition claims are made.

Alloy-like composition editing. The state is a six-element composition vector with fixed total mass. An edit increases one element and decreases another by the same amount. Intermediate feedback exposes estimated ultimate tensile strength (UTS), exact density, and trajectory signals such as recovery cost. The environment is a compositional optimization proxy with grounded components and controlled surrogate components; it is not a materials discovery benchmark. The diagnostic role is trajectory management: a locally attractive edit can improve one property while worsening the bounded recoverability signal or making successful completion require non-myopic recovery edits.

GB1 landscape editing. The state is a restricted sequence over selected GB1 positions. An edit changes one token. Intermediate feedback includes local additive estimates, but final success depends on an epistatic landscape grounded in the Fitness Landscape Inference for Proteins (FLIP) GB1 benchmark data (Dallago et al., 2021). The environment is a restricted deceptive-landscape stress test, not wet-lab protein engineering. The diagnostic role is exact recoverable construction: the agent must avoid local optima induced by misleading local evidence, and the finite landscape lets us audit whether failed trajectories crossed an exact recoverability boundary.

Interventions and diagnostics

We evaluate fixed prompting conditions under the same action-only interface: zero-shot, format-only few-shot, strategy-bearing few-shot, scaffold, generic self-check, and task-specific self-check. Prompt effects are interpreted as intervention-class diagnostics, not as model-specific prompt tuning. We also vary recent raw-history access ($M0$, $M1$, and $M3$, exposing zero, one, or three recent edits) to test whether trajectory context helps preserve feasible completion or instead entrenches poor local commitments.

We classify failures into surface failures and trajectory-level structural failures. Surface failures are malformed or locally invalid actions. Trajectory-level failures occur when the action interface is not the main bottleneck: the episode fails because the trajectory loses or degrades evidence for feasible completion, violates delayed constraints, finalizes prematurely, or falls into a local optimum. Recoverability drops and local-deception events provide mechanism evidence beyond post-hoc labels. In GB1 these diagnostics are checked against exact enumeration; in alloy-like editing they remain bounded-probe evidence.

4 Experimental Setup

The main controlled sweep uses `Qwen/Qwen2.5-7B-Instruct-Turbo` with deterministic hosted-model inference, fixed prompt templates, and seeds 0–49 per cell. The main prompt experiment evaluates all six prompts under state-only history ($M0$). The history ablation evaluates selected prompts under $M0$, $M1$, and $M3$. We report Wilson 95% confidence intervals for the headline success table and add an episode-level bootstrap audit, matched model–prompt cluster bootstrap, and fixed-effect logistic robustness check over the six-model raw logs in Appendix B.9. Intervals are descriptive diagnostics rather than hypothesis-testing claims.

The artifact also includes baseline calibration, alloy trajectory-state audit checks, GB1 operating-point validation, word-length ablation, six-model and 13-model cross-family checks, shallow wrapper pilots, non-LLM search pilots, raw per-episode logs, summary JSON files, and verification scripts. Cross-model results are used to test qualitative recurrence and model-conditional effects, not to establish a leaderboard or scaling law. Appendix A lists grounded components, controlled components, and claims not made for each environment.

Baseline hierarchies check that the environments are neither pure local optimization nor degenerate impossibility. For GB1, we additionally compute an exact state-budget recoverability table over the restricted landscape; for Alloy-like editing and GB1 we add lightweight non-LLM beam-search pilots in Appendix B.9. Random policies estimate chance success under the edit grammar; immediate-greedy policies test whether visible local improvement suffices; sequence-aware or privileged policies test whether non-trivial feasible completions exist. This hierarchy is essential for interpreting low LLM success as a trajectory-construction failure rather than a miscalibrated task.

5 Results

We begin with cross-model aggregates because they establish the surface-conditioned phenomenon across model families. Later prompt and history tables use the controlled Qwen sweep to interpret lightweight interventions under a fixed model and environment implementation.

5.1 A control-to-composition ladder

We first separate interface failure from composition failure. Table 1 aggregates the six-model cross-model sweep over three prompts and 50 seeds per cell. Word Ladder is the control case: local edit validity is only 36.1%, so final success partly reflects action surface failure. Alloy-like and GB1 editing shift the bottleneck. In alloy-like editing, the surface-clean rate is 83.8%, but success among surface-clean episodes is only 16.0%.

Table 1: Surface-conditioned cross-model evidence. Word Ladder is the action-interface control; Alloy-like and GB1 provide the main valid-but-unsuccessful composition evidence. Surface-clean is the episode-level rate of trajectories with no malformed or locally invalid action. Local valid is the action-level rate of accepted local edits, and Success | clean is success among surface-clean episodes. Exact counts are reported in Appendix B.9.

Environment	Local valid	Surface-clean	Success | clean
Word Ladder	36.1%	24.1%	75.1%
Alloy-like	67.6%	83.8%	16.0%
GB1	92.7%	84.4%	6.8%

Table 2: Trajectory diagnostics over an additional nine-model diagnostic-scope subset, separate from the six-model exact GB1 raw-log audit in Table 3. Recoverability-drop and local-deception rates are episode-level percentages. Finalize rate counts explicit `FINALIZE` actions; Word Ladder success may instead occur by reaching the target word before budget termination. For Alloy-like editing, a recoverability drop is a bounded-probe score decrease, not beam-certified binary infeasibility.

Environment	Recoverability drop	Local deception	Finalize rate
Word Ladder	67.9%	26.0%	0.0%
Alloy-like	85.7%	63.6%	79.7%
GB1	41.3%	43.5%	96.1%

In GB1, the surface-clean rate is 84.4%, local edit validity is 92.7%, and surface-clean success is only 6.8%. These models often satisfy the local edit interface while failing to construct a final artifact that meets the delayed objective.

This is the main separation the suite is designed to expose. The claim is not that all three environments instantiate the same failure mode. Instead, the matched protocol reveals a ladder of bottlenecks: Word Ladder diagnoses the surface interface, Alloy-like editing diagnoses delayed trade-off and trajectory management, and GB1 diagnoses exact recoverability loss under a deceptive landscape.

5.2 Trajectory-level failures dominate surface failures

Failure composition confirms that the residual bottleneck changes across the ladder. Word Ladder contains many surface failures and is dominated by BUDGET-UNAWARE continuation when it does produce valid moves. In the two composition environments, trajectory-level structural events dominate the event mix: 97.5% of surface-or-structural events in alloy-like editing and 85.3% in GB1. Alloy-like editing is dominated by OBJECTIVE-TRADEOFF failure, where local improvements worsen the strength–density trade-off or increase the need for explicit recovery search. GB1 is dominated by LOCAL-OPTIMUM TRAP, where locally attractive mutations lead away from globally successful combinations.

These labels should not be read as a new taxonomy of all LLM planning or search errors. They are deterministic diagnostics for this suite. Their value is that they align with the intended bottleneck of each environment: reachability and budget slack, delayed multi-objective feasibility, and deceptive local search.

5.3 Recoverability diagnostics locate the mechanism

We next ask where the trajectory fails. A recoverability drop occurs when an edit reduces the environment-specific estimate that the current partial solution still admits a feasible completion under the remaining budget. A local deception event occurs when a visible proxy improves while recoverability worsens. These probes are post-hoc diagnostics; they are not model-facing privileged signals.

The 0.0% Word Ladder finalize rate should not be read as absence of delayed evaluation. Word Ladder episodes terminate when the target is reached, the budget is exhausted, a dead end is reached, or a valid explicit `FINALIZE` is issued at the target; the reported finalize rate counts only the last case. Thus Word Ladder keeps the same terminal-evaluation interface while using target reach as an automatic successful termination.

Table 3 turns the mechanism claim into a measured decomposition. Word Ladder supplies the reachability-slack control. GB1 is the primary exact case: memoized enumeration confirms that the probe's binary recoverability labels match the full restricted state-budget table, and the observed 6.8% surface-clean success rate factors into a 51.4% boundary-loss term and a 14.1% residual conditional success term.

Alloy-like editing is deliberately weaker but complementary, and its audits answer two different questions. The bounded beam audit asks whether failed surface-clean trajectories can be certified impossible by a visible-state completion search; it finds no beam-certified losses, so we

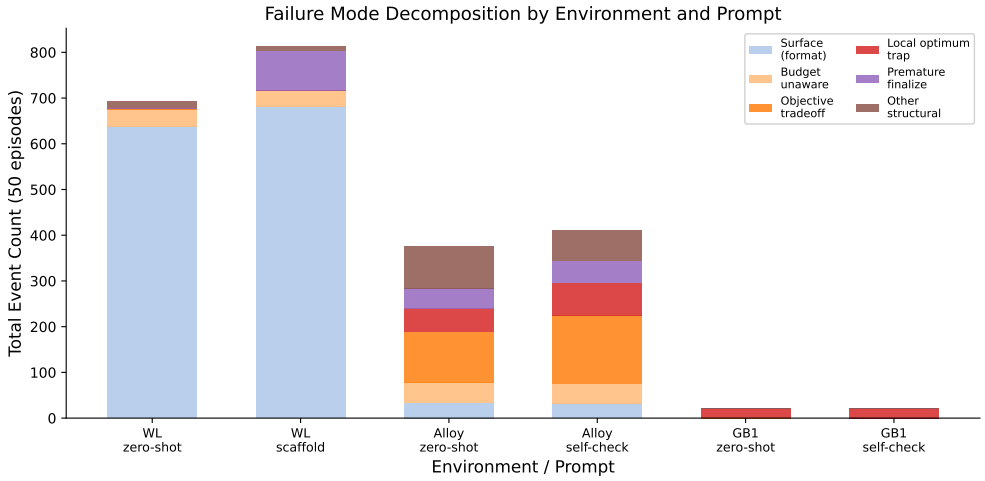

Figure 2: Failure composition across environments. Word Ladder acts as an interface-and-budget control. Alloy-like and GB1 editing show many trajectory-level structural failures even when surface failures are limited, indicating that their bottleneck is delayed construction rather than malformed action emission.

Table 3: Exact GB1 recoverability validation. The state-budget table enumerates all 625 restricted GB1 variants at budgets 0–6 using memoized dynamic programming over the finite edit graph and remaining-budget states, rather than reusing the online bounded probe. The trajectory audit uses the six-model GB1 raw logs from the cross-model panel and instantiates the surface-clean success decomposition in Section 3. Exact decomposition counts are reported in Appendix B.8.

Check	Quantity	Result
Exact state-budget table	pairs	4,375
Probe/exact mismatches	count	0
Raw GB1 audit	episodes	900
Boundary crossing rate	clean episodes	51.4%
Non-crossing success rate	episodes	14.1%
Loss before finalization	clean failures	55.2%
Exact local-deception rate	episodes	59.1%

should not treat bounded-probe drops as exact infeasibility. The implemented-state exact audits instead track the rounded simulator state, including path-dependent recovery cost, and show that many logged trajectories cross an implemented-state boundary. A broad cap-aware implemented-state audit labels 89.2% of visited states, agrees with the bounded probe on 99.5% of known states, and finds implemented-state exact loss for 97.0% of surface-clean failures. A stricter high-budget slice over visited states with remaining budget at most four labels 99.4% of states, agrees with the bounded probe on 98.0% of known states, and finds implemented-state exact loss for 95.9% of surface-clean failures. Alloy therefore remains mixed bounded evidence over the implemented rounded edit graph rather than a GB1-style continuous-state exact claim: some failures look recoverable under visible search, while many logged trajectories also show rounded-state boundary loss.

5.4 Recoverability measurements make intervention targets testable

The practical use of the measurement protocol is to turn trajectory labels into candidate intervention hypotheses, not only to report that an episode failed. Table 4 summarizes this mapping. The same final success rate can hide action-interface errors, poor stopping, local deception, or recoverable-but-not-found search failure; the trajectory diagnostics separate these cases. When we use prediction language, it is operational: the measurements say which repair class should be tested next, not that any wrapper is an optimal planner.

Cost-accounted wrapper checks partially test these hypotheses. In the open-family shallow-wrapper checkpoint (Appendix C.13, Table 15), alloy-like editing improves from 6.7% greedy success (8.0 calls/episode) to 22.0% under self-consistency (39.6 calls/episode), Word Ladder improves modestly (48.7% to 54.7%), and GB1 stays flat or worse (20.7% to 18.0%). A depth-3 visible-state beam leaves GB1 at 8–18% success, while switching beam scoring from `text_visible` to a privileged scorer raises alloy-like success from 11.7% to 21.7% without monotonically improving Word Ladder or GB1. The MCTS-style

Table 4: Recoverability measurements as intervention hypotheses. The rows are not a universal planner taxonomy; they state how trajectory labels change what should be tried next under this edit protocol.

Pattern	Hidden issue	Intervention target	Evidence in this paper
Low local validity	Interface or schema failure	Parser, examples, constrained decoder	Word Ladder control; retry effects
Surface-clean, low success	Valid actions miss delayed objective	Lookahead, rollback, search	Alloy wrapper/privileged-MCTS gains; GB1 visible-search limits
Recoverability drop/deception	Options consumed or proxy misleads	Guard, backtracking, uncertainty	GB1 exact loss; Alloy bounded drops
Finalize/history failure	Stopping or memory pathology	Pre-finalize verifier; summaries	Finalize and history ablations

boundary check is sharper: visible-proxy MCTS solves 74.0% of GB1 but only 6.0% of Alloy seeds, whereas privileged terminal-objective MCTS solves all GB1 and 82.0% of Alloy seeds. Calls per episode and available token accounting are reported in the released summaries.

5.5 Format examples are not a robust repair

Table 5 reports the main controlled prompt sweep. If weak performance were primarily a schema-compliance problem, format-only few-shot demonstrations should improve broadly. In this sweep they do not: format-only prompting is below zero-shot in Word Ladder, alloy-like editing, and GB1. The best intervention also differs by environment. Word Ladder benefits most from a scaffold, while alloy-like and GB1 editing benefit most from task-specific self-checking.

Thus prompts act as lightweight policy-shaping interventions, not uniform repairs: scaffolds help the reachability-control case, task-specific checks help the deceptive-landscape case, and alloy-like trade-offs remain difficult.

5.6 Raw history is conditional, not memory magic

Appendix F shows that recent raw edit history is not uniformly useful. In the controlled Qwen ablation, Alloy-like zero-shot improves from 12% under state-only history ($M0$) to 25% with three recent edits ($M3$), but Word Ladder scaffold drops from 30% to 10%, and the selected GB1 self-check row drops from 35% to 5%.

5.7 Cross-model checks bound the claim

The six-model sanity check and 13-model open-family panel show qualitative recurrence without supporting a leaderboard or scale law. Across the six-model grid, format-only exemplars remain inconsistent and structural events dominate surface events in alloy-like and GB1 editing. The larger structural-prompt panel in Appendix E shows the same scope qualification: Llama-3.1-8B and Qwen3-14B tie for the top Word Ladder result, Llama-3.3-70B leads alloy-like editing, and Ministral-8B leads GB1. No model is uniformly best across the suite.

Stability and sensitivity checks in Appendix C.13 show that scorer, decoding, hosted-model API perturbations, memory, prompt, and surrogate choices can move cell-level rates. We therefore state recurrence claims qualitatively: tested models often show recoverability degradation despite locally admissible edits, but exact rates and best interventions remain model-conditional.

5.8 Calibration rules out degenerate interpretations

Calibration checks support the operating points. Word Ladder requires reachability and budget slack beyond repeated valid edits; alloy-like editing is not pure local-property maximization because strength gains can worsen density or recovery cost; and GB1 is not explained by additive local improvement because additive greedy baselines fail across tested thresholds. Appendix B.9 adds bootstrap, paired cluster-bootstrap, fixed-effect logistic, and non-LLM beam checks; together they preserve the separation between high Word Ladder surface-clean success and low Alloy-like or GB1 surface-clean success, supporting trajectory-construction failure rather than degenerate task impossibility.

Table 5: Main prompt-intervention results at state-only history ($M0$). Values are success rates (%) with Wilson 95% confidence intervals over 50 episodes per cell. Format-only examples do not reliably improve over zero-shot, and the strongest intervention differs by environment.

Environment	Zero-shot	Few-shot (format)	Few-shot (strategy)	Scaffold	Self-check (generic)	Self-check (task-specific)
Word Ladder (4-letter)	18 [10,31]	8 [3,19]	18 [10,31]	**30 [19,44]**	6 [2,16]	18 [10,31]
Alloy-like	12 [6,24]	10 [4,21]	2 [0,10]	14 [7,26]	6 [2,16]	**16 [8,29]**
GB1 landscape	24 [14,37]	20 [11,33]	18 [10,31]	12 [6,24]	26 [16,40]	**38 [26,52]**

6 Discussion

Taken together, the results argue for treating recoverability as a first-class measurement target in sequential agent evaluation. The three environments do not instantiate one generic planning failure; they expose a progression of increasingly structural bottlenecks under a shared edit interface. Word Ladder anchors the action-surface regime, where low local validity (36.1%) makes terminal failure hard to interpret as trajectory mismanagement. Alloy-like editing and GB1 shift the question: both are often surface-clean, yet success among clean trajectories remains 16.0% and 6.8%. The gap between local admissibility and delayed success is therefore not a formatting artifact. It is the phenomenon the protocol is designed to measure.

Recoverability turns that gap from a terminal statistic into a trajectory-level diagnostic. In GB1, exact enumeration over the restricted 625-variant landscape separates failures that cross a recoverability boundary from failures that occur despite retained feasibility. This distinction is unavailable from success rate alone, but it is essential for interpreting what the model failed to preserve. Alloy-like editing provides a different, intentionally bounded form of evidence: recoverability probes, rounded-state audits, and visible-state beam checks point to delayed trade-off burden without asserting continuous-state exactness. The evidential status is therefore tiered by design: Word Ladder is a control, Alloy-like editing is a bounded mechanism probe, and GB1 is the exact decomposition case.

The intervention results reinforce the same interpretation. Format examples, raw history, wrapper sampling, and privileged search are not interchangeable ways to raise a score; they test different hypotheses about where the trajectory breaks. Their uneven effects are informative: history can help or hurt, sampling helps Alloy-like editing more than GB1, and privileged objective information matters only when it aligns with the measured bottleneck. Table 4 should therefore be read as a mechanism-to-repair map. It connects observed trajectory evidence to plausible intervention classes, rather than treating all failures as either prompt defects or generic search deficits.

The broader implication is a reporting standard for constructive agents. Final success should be accompanied by local validity, surface-clean success, recoverability loss, local deception, finalize behavior, dominant failure class, intervention cost, and the evidential status of any oracle or bounded probe. These measurements do not make the controlled proxies into deployment benchmarks for materials or proteins. They make the scope of the claim explicit: valid local edits can stop composing into feasible artifacts, and evaluation should reveal when, how, and under which repair assumptions that breakdown occurs.

7 Conclusion

We introduced a recoverability-centered measurement protocol for sequential construction over partial structured outputs. A matched edit substrate separates action-surface failure, delayed trade-off burden, and exact recoverability loss under deceptive local evidence. Across the three environments, the same local edit interface exposes different failure regimes: malformed or invalid actions, surface-clean trajectories that miss delayed trade-offs, and exact boundary loss in a finite deceptive landscape. Recoverability is therefore a measurable, auditable, and intervention-informative reporting axis for when valid local edits cease to preserve feasible completion. Reporting this axis alongside final success makes sequential evaluation less about whether an agent can take a valid next step and more about whether its accepted steps continue to leave a solution possible.

8 Limitations

Constructive abstraction versus full realism. Our environments intentionally abstract away much of the complexity of real scientific design workflows. Alloy development involves thermodynamics, kinetics, processing conditions, and microstructure, while real biological sequence design involves richer structural, functional, and experimental constraints than the restricted GB1 landscape proxy represented here. We therefore position the benchmark as an evaluation of *constructive editing structure*, not of full domain realism.

Surrogate–reality gap. Alloy-like composition editing and GB1 landscape editing rely on surrogate or semi-synthetic scorers rather than full simulation or wet-lab validation. This is necessary for controlled large-scale evaluation, but it limits claims about scientific deployment.

Restricted edit grammars. Although the environments do not reduce to menu selection, they still use constrained edit grammars to avoid confounding constructive trajectory management with arbitrary free-form generation. This improves interpretability, but underrepresents fully open design spaces.

Scale of edit horizons. The edit budgets are intentionally modest to keep evaluation reproducible under hosted-model inference. These budgets are sufficient to induce path dependence, but they remain shorter than many real design workflows.

Recoverability validation. For GB1, we validate the recoverability probe against exact enumeration over the restricted 625-variant landscape and budgets 0–6. Alloy-like editing remains a bounded-probe analysis: the fixed delta grid makes exhaustive or large-beam feasibility auditing plausible, and we include a trajectory-state beam audit in Appendix B.9, but we do not claim full precision/recall against the continuous alloy state space. In that audit, no audited surface-clean Alloy failures show beam-certified loss, so Alloy recoverability claims should be read as diagnostic evidence of trade-off and lookahead burden. Finite-state exact audits over the implemented rounded Alloy edit graph now include an all-visited cap-aware audit and a higher-cap slice over states with remaining budget at most four, but this is still not a full continuous-state feasibility proof. The obstacle is not only state count: recovery cost is a path-dependent rounded latent variable, so an exact all-budget Alloy table must track composition, phase, remaining budget, and recovery-cost state rather than composition alone.

Planner and statistical baselines. The added non-LLM beam pilots rule out degenerate impossibility for the audited starts, and the MCTS-style boundary check separates visible-proxy search from privileged terminal-objective search. These are still boundary checks, not a full planning study. Stronger baselines would include learned value functions, explicit rollback/backtracking policies, and richer information-assumption sweeps. Likewise, bootstrap intervals quantify episode-level uncertainty, but the paired model–prompt cluster bootstrap and fixed-effect logistic adjustment are still robustness checks rather than a full mixed-effects model over all crossed sources of variation.

Search-wrapper scope. Our main experiments evaluate action-only editing policies under lightweight prompt and history interventions, and Appendix C.12 adds a matched shallow-wrapper checkpoint with greedy sampling, self-consistency, depth-1 beam search, and loop avoidance under a `text_visible` scorer, plus a depth-3 visible-state beam stress test and a non-LLM MCTS-style boundary check. The MCTS check uses an explicit action enumerator, and its privileged mode uses terminal-objective information, so it is not an attainable LLM policy. We still do not claim to characterize learned value functions, explicit rollback/backtracking, or tool-assisted self-refinement. The wrapper checkpoint should be read as a cost-accounted search stress test, not as a privileged sensitivity reference for search-aided LLM editing policies.

Exemplar sensitivity and model scope. The alloy robustness ablation shows that strategy-bearing demonstrations can be highly sensitive to the exact trajectory shown. The main controlled study uses one model family, which improves causal interpretation but limits generalization across providers and alignment regimes. The cross-model sanity check, the 13-model open-family panel (§5.7, Appendix C.13), and the shallow-wrapper checkpoint broaden this evidence, but they remain benchmark probes rather than a comprehensive leaderboard — the model grid is intentionally unbalanced across families and sizes, and the panel is meant to demonstrate

discrimination, stability, and sensitivity rather than to estimate a scaling law. Accordingly, our claims concern controlled constructive trajectory-management burdens rather than a universal account of all LLMs, full materials or biological design, or biological mechanism modeling.

9 Ethical Considerations

This work uses public lexical, alloy-property, and protein fitness-landscape data to construct controlled evaluation environments. It does not involve human subjects, private user data, annotation labor, or deployment decisions. The scientific-design-inspired tasks are explicitly scoped as diagnostic proxies: the artifact contains no wet-lab protocol, synthesis instruction, DNA-design workflow, or biosynthetic design guidance. The main risk is overinterpreting proxy success as deployable scientific capability, which we mitigate by separating grounded and controlled components, reporting claim boundaries, and releasing raw logs and verifier scripts for audit. Hosted-model experiments use documented API calls for evaluation only; the submission-facing artifact is checked for secret leakage and does not require reviewers to run hosted calls.

References

Christopher K. H. Borg, Carolina Frey, Jasper Moh, Tresa M. Pollock, Stéphane Gorsse, Daniel B. Miracle, Oleg N. Senkov, Bryce Meredig, and James E. Saal. 2020. Expanded dataset of mechanical properties and observed phases of multi-principal element alloys. *Scientific Data*, 7(1):430.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, and Lilian Weng. 2025. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*.

Christian Dallago, Jody Mou, Kadina Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Emmanuele Papa, Sabina Shakhimardanova, Alex Cornett, Johannes Welbl, Aleksandr Shtedritski, Boris Burkov, Ziqing Yang, Axel X. Lu, Anastasiia Kozlova, Sungroh Yoon, Alaa Nawy, Sarah Min, Steven E. Brenner, and 18 others. 2021. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems*.

Janardhan Rao Doppa, Alan Fern, and Prasad Tadepalli. 2014. Structured prediction via output space search: The search-based structured prediction framework. *Journal of Machine Learning Research*, 15:1317–1350.

Yuanqi Du, Yingheng Wang, Yining Huang, Jianan Canal Li, Yanqiao Zhu, Tian Xie, Chenru Duan, John M. Gregoire, and Carla P. Gomes. 2023. M^2hub: Unlocking the potential of machine learning for materials discovery. In *Advances in Neural Information Processing Systems*.

Gonzalo Gonzalez-Pumariega, Leong Su Yean, Neha Sunkara, and Sanjiban Choudhury. 2025. Robotouille: An asynchronous planning benchmark for LLM agents. In *International Conference on Learning Representations*.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems*, pages 11179–11189.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*.

Harsha Kokel, Michael Katz, Kavitha Srinivas, and Shirin Sohrabi. 2025. ACPBench: Reasoning about action, change, and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26559–26568.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. Agentbench: Evaluating LLMs as agents. In *International Conference on Learning Representations*.

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

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods*

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5054–5065.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wołczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. 2025. BALROG: Benchmarking agentic LLM and VLM reasoning on games. In *International Conference on Learning Representations*.

Benjamín Sánchez-Lengeling and Alan Aspuru-Guzik. 2018. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365.

Mahakaran Sandhu, John Z. Chen, Dana S. Matthews, Matthew A. Spence, Sacha B. Pulsford, Barnabas Gall, Joe A. Kaczmarski, James Nichols, Nobuhiko Tokuriki, and Colin J. Jackson. 2025. Computational and experimental exploration of protein fitness landscapes: Navigating smooth and rugged terrains. *Biochemistry*, 64(8):1673–1684.

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

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

Weiwei Sun, Shengyu Feng, Shanda Li, and Yiming Yang. 2026. CO-Bench: Benchmarking language model agents in algorithm search for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33126–33134.

Jia Wang, Yingxue Wang, and Yanan Chen. 2022a. Inverse design of materials by machine learning. *Materials*, 15(5):1811.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022b. Scienceworld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11279–11298.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*.

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

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. 2024. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*.

Appendix

This appendix provides supporting formalization, environment-construction details, detailed results tables, representative failure patterns, and abbreviated prompt templates.

A Environment Grounding and Claim Boundaries

Table 6: Grounded components, controlled components, and claim boundaries for the three environments. The suite tests constructive editing structure, not full realism in lexical creativity, materials discovery, or biological engineering.

Environment	Grounded component	Controlled component	Not claimed
Word Ladder	ENABLE lexicon and Hamming-1 word graph	sampled start–target pairs, edit budget, no shortest-path hint	natural-language generation or lexical creativity
Alloy-like composition editing	Multi-Principal Element Alloy (MPEA) measured-property data and exact density calculation	heavy-element scaling, recovery cost, thresholds, observation noise	real materials discovery or process-aware alloy development
GB1 landscape editing	public Fitness Landscape Inference for Proteins (FLIP) GB1 fitness landscape over measured variants	four editable sites, noisy additive observations, stability gate	wet-lab protein engineering or biosynthetic design protocol

B Methodological Formalization

This appendix states the definitions used in the main text. The goal is to make clear which quantities are model-facing task variables and which are post-hoc diagnostics. We use terminology from sequential decision making, structured prediction as search, and local search, while using *constructive structure editing* only as an operational label for the matched task family.

B.1 Constructive Editing Task

For environment e, a constructive-editing task is

$$\mathcal{T}_e = (\mathcal{S}_e, \mathcal{A}_e, P_e, O_e, G_e, H_e), \tag{2}$$

where $\mathcal{S}_e$ is the partial-solution state space, $\mathcal{A}_e$ is the set of local edits plus `FINALIZE`, P_e is the edit transition function, O_e is the observation function, G_e is the delayed global evaluator, and H_e is the edit budget. At step t, the model observes

$$o_t = O_e(x_t, \tau_{<t}), \tag{3}$$

where $x_t \in \mathcal{S}_e$ is the current partial structured output and $\tau_{<t}$ is the previous edit trajectory. The model emits either a local edit a_t or `FINALIZE`. A local edit updates the state by

$$x_{t+1} = P_e(x_t, a_t), \tag{4}$$

while `FINALIZE` triggers delayed evaluation

$$y = G_e(\tau_{\leq t}, x_t). \tag{5}$$

We distinguish action-interface properties from trajectory properties:

- **Well-formed action**: the string matches the expected action grammar.
- **Locally admissible edit**: the parsed edit is accepted by the environment from the current state.
- **Surface-clean episode**: every emitted action is well formed and locally admissible.

- **Delayed-objective success**: final evaluation satisfies the environment success predicate $S(\tau, x_T) = 1$.
- **Feasible completion**: the current partial solution still admits at least one successful completion under the remaining edit budget.
- **Recoverability diagnostic**: an environment-specific measured or estimated label for feasible completion, used only in post-hoc analysis.

The central diagnostic question is whether a locally admissible edit preserves diagnostic recoverability toward feasible completion. This is weaker and more precise than a general claim about LLM planning: it asks whether a chosen edit keeps the remaining construction problem solvable under the benchmark's fixed constraints.

B.2 Why This Is Sequential Decision Making

The tasks can be viewed as finite-horizon, partially observed sequential decision problems: the environment has a state, transition function, observation function, action grammar, horizon, and terminal reward, while the model observes only the rendered task state and local feedback. They are also instances of structured prediction as search over partial outputs. Our terminology is narrower than either framework: it selects tasks where every action edits an auditable partial artifact and where delayed success can be decomposed into interface and trajectory-state diagnostics. A task is included only if it satisfies four properties.

Local editability. The model must transform a partial solution through local actions rather than select a complete candidate in one step.

Partial-state semantics. Intermediate states are meaningful artifacts: a word on a path, a composition vector, or a sequence variant.

Path dependence. Earlier edits change the set or quality of feasible future completions. The same locally admissible edit can be useful or harmful depending on the current trajectory and remaining budget.

Delayed global evaluation. The environment evaluates success only at terminal evaluation: `FINALIZE`, budget termination, dead-end termination, or the environment-specific target-reach termination used by Word Ladder. Local observations may be informative, but they do not certify final success except when they trigger such a terminal state.

B.3 Environment-Specific Instantiations

Lexical editing. The partial solution is a word path in a Hamming graph. A locally admissible edit changes exactly one character, yields a lexicon word, and avoids revisits. Feasible completion corresponds to retaining a path to the target within the remaining edit budget.

Alloy-like composition editing. The partial solution is a fixed-mass composition vector. A locally admissible edit transfers mass between elements. Feasible completion corresponds to remaining able to satisfy delayed thresholds on ultimate tensile strength (UTS) and density without excessive recovery cost. This is a controlled compositional proxy, not a real materials discovery claim.

GB1 landscape editing. The partial solution is a restricted sequence variant. A locally admissible edit changes one selected position. Feasible completion corresponds to remaining able to reach a successful variant on the restricted epistatic landscape. This is a sequence-landscape proxy, not wet-lab protein engineering or DNA design.

B.4 Prompted Policy and History Conditions

A model under prompt condition p and history condition k induces a policy $\pi_m^{(p,k)}(a_t \mid o_t)$. We abbreviate success rate as SR in appendix formulas and detailed result tables, and report it as

$$\mathrm{SR}(e, m, p, k) = \mathbb{E}_{\tau \sim \pi_m^{(p,k)}, \mathcal{T}_e} \left[S(\tau, x_T) \right]. \tag{6}$$

Prompt conditions are fixed intervention classes. They are not tuned per model. History conditions expose raw recent trajectory information: $M0$ is state-only history, while $M1$ and $M3$ expose one or three recent edits. They should not be interpreted as a general memory-system benchmark.

B.5 Prompt and History Effects

For a prompt intervention p against zero-shot p_0, we define the descriptive prompt effect

$$\Delta_{\text{prompt}}(e, p) = \text{SR}(e, m, p, M0) - \text{SR}(e, m, p_0, M0). \tag{7}$$

For history condition k, we define

$$\Delta_{\text{hist}}(e, p, k) = \text{SR}(e, m, p, k) - \text{SR}(e, m, p, M0). \tag{8}$$

These are descriptive diagnostics. Unless explicitly supported by uncertainty analysis, we interpret them as qualitative patterns rather than statistically separated treatment effects.

B.6 Failure Events

We group failure events by what they diagnose.

Surface failures. These are action-interface failures: malformed outputs, unparsable actions, or locally inadmissible edits.

Edit-process failures. These are valid-interface behaviors that prevent appropriate termination or state progression, such as premature finalization or repeated continuation after the remaining budget is no longer sufficient.

Trajectory-level structural failures. These are failures in which the episode is not primarily blocked by the action interface. Examples include budget-unaware continuation in Word Ladder, objective-tradeoff collapse in alloy-like editing, recovery-cost growth, and local-optimum trapping in GB1. The phrase *structural* is used only in this diagnostic sense: the trajectory shows recoverability degradation or violates a delayed global constraint despite locally admissible actions.

B.7 Recoverability and Local Deception

Let $R_e(x_t, b_t)$ be an environment-specific diagnostic probe for whether the current state x_t with remaining budget b_t admits a feasible completion. A **recoverability drop** occurs when an edit reduces this probe, including a diagnostic transition from recoverable to unrecoverable. A **local deception event** occurs when a visible local proxy improves while R_e worsens. These diagnostics are used only to analyze stored trajectories. They are not model-facing privileged signals, they are not claimed to be universal measures of planning difficulty, and they should not be interpreted as exact feasibility labels unless separately validated for the environment.

B.8 Recoverability Probe Scope

The reported recoverability probes are environment-specific diagnostics. In Word Ladder, the graph structure makes reachability under the remaining budget closest to exact. In GB1, the restricted landscape is small enough for exact feasible-completion enumeration. We enumerate all 625 variants over budgets 0–6 with an independently implemented memoized search over the finite edit graph and remaining-budget states, then compare the environment probe against those exact labels; agreement is 100% with zero false positives and zero false negatives. We also audit the six-model GB1 raw logs. In the decomposition over all surface-clean episodes, 391/760 cross an exact recoverability boundary and 52/369 residual non-crossing episodes succeed; equivalently, 391 of 708 surface-clean failures cross the exact boundary before finalization. In alloy-like editing, the discrete edit grid makes exhaustive or large-beam feasibility analysis plausible under fixed delta sets, but the current results still use a bounded diagnostic probe. Thus GB1 recoverability is exact in the reported restricted landscape, while alloy recoverability remains diagnostic: an Alloy recoverability drop denotes bounded-probe degradation, not a beam-certified transition to true infeasibility.

Table 7: GB1 exact recoverability validation used to support the mechanism analysis.

Audit	Unit	Result
Restricted landscape size	variants	625
Budgets enumerated	edit budgets	0–6
Probe/exact agreement	variant-budget pairs	4,375/4,375
False positives / negatives	variant-budget pairs	0 / 0
Raw-log probe/exact agreement	diagnostic steps	4,779/4,779
Surface-clean GB1 failures with exact loss	episodes	391/708
Exact local-deception episodes	episodes	532/900

B.9 Soundness Audits: Bootstrap, Alloy Beam Audit, and Non-LLM Search

We add five lightweight audits aimed at reviewer concerns about diagnostic soundness. First, an episode-level bootstrap over the six-model raw logs confirms that the headline pattern is not a single-cell artifact: surface-clean success remains high for Word Ladder but low for Alloy and GB1. Second, a paired cluster bootstrap over matched model–prompt cells checks that the ordering is not driven by one model family or one prompt class. Third, a grouped-binomial fixed-effect logistic check adjusts for model and prompt-class indicators and preserves the environment ordering. Fourth, for Alloy we audit unique trajectory states visited in the raw logs with a depth-capped beam over the same fixed edit grid. This is not a full continuous-state exact feasibility proof; it is a trajectory-state check that the bounded probe is not arbitrary on visited states and that many failed surface-clean trajectories remain recoverable by explicit search. Fifth, we run non-LLM beam-search pilots under the same edit grammar. Beam-found completions are accepted only when they satisfy the same environment terminal predicate used for LLM episodes. These pilots test whether feasible completions exist under simple search, rather than whether LLMs can discover them with one action per call.

B.10 Baseline Hierarchy

The baseline hierarchy checks operating-point validity rather than ranking agents. Random policies estimate chance success under the grammar. Immediate greedy policies test whether visible local improvement suffices. Sequence-aware or privileged diagnostic policies test whether non-trivial feasible completions exist. A useful environment should not be solved by greedy local optimization, but should remain solvable by a policy with access to the relevant trajectory-level structure.

C Environment Construction Details

This appendix describes how the three evaluation environments were constructed and calibrated. Our goal was not to maximize domain realism, but to instantiate a shared constructive-editing protocol in which local edits, path dependence, and delayed global evaluation are all consequential. We document the design goals, instance construction procedures, operating-point selection, and filtering rules used in the final benchmark.

C.1 Design Goals

The environments were designed to satisfy five criteria:

1. **Irreducibility to one-shot generation.** Each environment requires a sequence of local edits rather than a single output, so that trajectory-level failures can emerge.

2. **Valid-but-suboptimal trajectories.** At every step the agent faces multiple valid actions, most of which lead to suboptimal or failed final states. Surface formatting correctness therefore does not predict task success.

3. **Baseline separability.** The operating regime was chosen so that random, observation-level greedy, and privileged diagnostic baselines occupy distinct performance tiers.

Table 8: Soundness audits added beyond final success rates. Bootstrap CIs use six-model raw logs. The Alloy beam audit is restricted to unique visited trajectory states in the smoke-scale audit: the 54/54 row is state-level beam/probe agreement, while the 0/633 row is an episode-level check that no audited surface-clean Alloy failure shows beam-certified loss. The all-visited implemented-state audit uses a 10,000-expansion cap per state. The high-budget exact slice uses a 250,000-expansion cap over visited states with remaining budget at most four. In both audits, capped states are reported as unknown. Non-LLM beam pilots are operating-point checks, not leaderboard results.

Audit	Unit	Result
Word Ladder surface-clean success	bootstrap 95% CI	75.1 [69.6, 80.2]
Alloy surface-clean success	bootstrap 95% CI	16.0 [13.3, 18.8]
GB1 surface-clean success	bootstrap 95% CI	6.8 [5.3, 8.7]
Word Ladder surface-clean episodes	episodes	217/900
Alloy surface-clean episodes	episodes	754/900
GB1 surface-clean episodes	episodes	760/900
Word Ladder–Alloy surface-clean gap	paired cluster bootstrap	+58.8 [43.5, 71.6] points
Word Ladder–GB1 surface-clean gap	paired cluster bootstrap	+69.7 [57.9, 80.3] points
Alloy–GB1 surface-clean gap	paired cluster bootstrap	+10.9 [4.2, 17.4] points
Adjusted surface-clean success	fixed-effect logit	75.0 / 15.7 / 6.6%
Alloy vs. Word Ladder adjusted odds ratio	fixed-effect logit	0.061 [0.024, 0.152]
GB1 vs. Word Ladder adjusted odds ratio	fixed-effect logit	0.023 [0.009, 0.061]
Alloy vs. GB1 adjusted odds ratio	fixed-effect logit	2.641 [1.052, 6.629]
Alloy beam/probe agreement	visited states	54/54
Alloy surface-clean failures with beam-certified loss	episodes	0/633
Alloy all-visited states exact-labeled	visited states	5477/6143
Alloy all-visited exact/probe agreement	known visited states	5449/5477
Alloy all-visited failures with exact loss	surface-clean failures	614/633
Alloy high-budget implemented states exact-labeled	budget $\leq$4 states	4166/4193
Alloy high-budget exact/probe agreement	known budget $\leq$4 states	4084/4166
Alloy failures with implemented exact loss	surface-clean failures	607/633
GB1 exact guard coverage	surface-clean failures	391/708
GB1 residual exact-recoverable failures	surface-clean failures	317/708
Alloy visible beam pilot	starts	10/10
Alloy privileged beam pilot	starts	10/10
GB1 visible beam pilot	starts	20/20
GB1 privileged beam pilot	starts	20/20

4. **Shared interaction protocol.** All three environments expose the same interface: a text observation, an `EDIT` / `FINALIZE` action grammar, per-step validity feedback, and a terminal success signal.

5. **Consequential terminal evaluation.** The globally evaluated quantity is resolved only at terminal evaluation (`FINALIZE`, budget exhaustion, dead end, or environment-specific success termination), forcing the agent to plan under incomplete information.

C.2 Shared Interaction Protocol

Every environment implements the following loop:

1. **Reset.** A seed-dependent starting state is sampled and an initial observation is rendered as natural-language text.

2. **Step.** The agent produces a raw text action. The environment parses it (`EDIT . . .` or `FINALIZE`), validates it against domain constraints, applies it if valid, and returns an updated observation together with any failure events.

3. **Termination.** The episode ends when the agent issues a valid `FINALIZE`, the step budget is exhausted, or a dead-end state is reached. Word Ladder additionally auto-terminates successfully when an accepted edit reaches the target word; its reported finalize rate therefore counts only explicit `FINALIZE` actions, not target-reach termination.

4. **Success criterion.** Evaluated at termination against the environment's hidden global objective.

The action grammar is environment-specific but always begins with `EDIT` followed by positional and value arguments. Invalid actions consume a step but do not alter the state. Table 9 summarizes the key dimensions of each environment.

Table 9: Environment summary. All environments share the same interaction protocol—text observation, `EDIT`/`FINALIZE` actions, local validity feedback, and delayed terminal evaluation— but differ in state representation, action grammar, and the dominant source of constructive difficulty. The table highlights why the benchmark is comparable across domains without collapsing them into the same task.

	Word Ladder	**Alloy-like**	**GB1 landscape**
Structure type	string (word)	composition vector	amino acid sequence
State dimension	4 characters	6 element percentages	4 residue positions
Action grammar	`EDIT pos letter`	`EDIT inc dec` δ	`EDIT pos aa`
Action space (approx.)	4×25	$30 \times \lvert\Delta\rvert$	4×4
Budget	$5 \times k_{\text{opt}}$	8 steps	6 steps
Evaluation instances	409 pairs	continuous (jittered)	625 variants
Observation noise	none	strength ± 30 MPa	fitness ± 0.20
Key difficulty	dead-end avoidance	strength–density tradeoff	epistasis trap

C.3 Word Ladder Instance Construction

Lexicon and graph. We use the Enable1 word list[1], a standard Scrabble tournament dictionary. After filtering to 4-letter words we obtain a vocabulary of 3903 words. We construct a Hamming-1 adjacency graph over this vocabulary ($|E| = 21\,002$); the largest connected component contains 3807 words (97.5%). The average degree is 9.9.

Pair selection. We sample 500 candidate start–target pairs from the largest connected component and compute breadth-first-search shortest-path lengths. We retain pairs with shortest-path distance $k_{\text{opt}} \in [3, 6]$, yielding **409 pairs**. The lower bound excludes trivial one- or two-step instances, and the upper bound excludes instances that require very long optimal paths.

Budget and observation. Each episode receives a budget of $\max(5 \times k_{\text{opt}},\ 15)$ steps. At each step the agent sees the target word, the current word with mismatch markers, the path taken so far, the remaining budget, and a list of valid replacement letters per position. The observation does *not* include shortest-path distance to the target. Reaching the target by an accepted edit automatically terminates the episode as success; an explicit `FINALIZE` action is also valid at the target, but is not required for Word Ladder success.

Word length justification. We select 4-letter words as the main setting based on a difficulty ablation (Appendix K). Three-letter words are easier overall, while five-letter words are substantially sparser and more dead-end prone. Four-letter instances provide the most informative operating regime.

C.4 Alloy-like Composition Editing Environment Construction

Composition space. The agent edits a six-element alloy composition over Fe, Ni, Cr, Co, Mn, and Mo, subject to a simplex constraint ($\sum_i c_i = 100\%$) and per-element bounds.

Action grammar. Each `EDIT inc dec` δ transfers δ atomic percent from one element to another. Allowed δ values depend on the phase: coarse ($\delta \in \{10, 15\}$) during the first half of the budget, and fine ($\delta \in \{2, 5\}$) during the second half. This creates order sensitivity.

Ultimate-tensile-strength surrogate. We train a Gradient Boosting Regressor on 539 measured alloys from the Multi-Principal Element Alloy (MPEA) mechanical properties dataset (Borg et al., 2020). The surrogate predicts ultimate tensile strength (UTS) from the six-element composition vector.

[1] https://norvig.com/ngrams/enable1.txt

Heavy-element strength scaling. To create a genuine strength–density tradeoff, we apply a multiplicative scaling factor to the base UTS prediction:

$$\mathrm{UTS}_{\mathrm{eff}} = \mathrm{UTS}_{\mathrm{base}} \cdot \left[\alpha + (1-\alpha) \cdot \min\left(1,\ \frac{h}{h_{\mathrm{thresh}}}\right)\right], \tag{9}$$

where $h = 1.0 \cdot c_{\mathrm{Mo}} + 0.6 \cdot c_{\mathrm{Co}} + 0.3 \cdot c_{\mathrm{Ni}}$, $\alpha = 0.30$, and $h_{\mathrm{thresh}} = 18.0$. Without sufficient heavy elements, predicted strength is capped well below the target, making it impossible to reach the strength threshold without including density-increasing elements.

Mechanism diagnostic. To check that the alloy-like proxy actually induces the intended tradeoff, we analyze all valid one-step edits from 100 seeded starting compositions. Across 3,833 valid first edits, 62.1% improve one objective while worsening the other. Among strength-improving edits, 78.9% increase density; among density-improving edits, 79.5% reduce strength. A 20-seed deterministic baseline sensitivity check further shows that privileged sequence-aware success remains high in the default setting (95%), rises under the easy combined setting (100%), and drops under the hard combined setting (55%). This supports the interpretation that the environment is neither a pure local-strength maximization task nor an arbitrary impossible regime, while the LLM sensitivity evidence remains limited to the targeted OpenAI sweeps in Appendix C.12.

Density and recovery cost. Density is computed as the composition-weighted average of elemental densities. To create path dependence, we also track a recovery cost that accumulates when the composition moves away from a known-good region. If recovery cost exceeds 3.8 at `FINALIZE`, the episode fails even if strength and density targets are met.

Starting composition and observation. The default composition is Fe:25, Ni:15, Cr:20, Co:10, Mn:20, Mo:10 (at.%). For seed-dependent variation, each element is perturbed by $\pm 8\%$ and renormalized. The agent observes the current composition, a noisy strength estimate (± 30 MPa Gaussian), exact density, a qualitative path-efficiency indicator, the current phase, and recent edit results.

Success criterion. An episode succeeds if and only if all three conditions hold at `FINALIZE`: $\mathrm{UTS}_{\mathrm{eff}} \geq 2{,}000$ MPa, density ≤ 8.20 g/cm^3, and recovery cost < 3.8.

C.5 GB1 Landscape Editing Environment Construction

Biological basis. We base this environment on the GB1 fitness landscape from the Fitness Landscape Inference for Proteins (FLIP) benchmark (Dallago et al., 2021), which provides experimentally measured binding fitness for a large set of GB1 variants.

Editable positions. We select four positions with strong epistatic interactions: V39, D40, G41, and V54. Each position admits five residues, yielding a combinatorial space of $5^4 = 625$ variants. The wild-type sequence is VDGV.

True fitness and the epistasis trap. True fitness is looked up from a precomputed table derived from the FLIP data. Only 44 of 625 variants (7.0%) have fitness ≥ 5.0. The environment is designed so that single-site fitness estimates are systematically misleading about combinatorial fitness: top variants require mutations at position 2, despite position 2 appearing harmful under additive local estimates. This creates the intended local-optimum trap.

We quantify this mismatch directly over the full 625-variant landscape. The additive estimate has weak association with true epistatic fitness ($r = 0.20$, Spearman $\rho = 0.20$). The top 5 and top 10 variants under the additive estimate have zero overlap with the true top 5 and top 10 variants; overlap remains only 8% at top 25 and 16% at top 50. Among variants meeting the success threshold, 40 of 44 (90.9%) require a mutation at the destabilizing position 2. These diagnostics support the intended interpretation that local additive evidence is not merely noisy but structurally misaligned with the delayed epistatic objective.

Stability mechanic. To create path dependence, each mutation carries a stability delta (Table 10). Position 2 mutations are especially destabilizing. If stability drops below a gate threshold of 0.55, further destabilizing mutations are blocked; if stability falls below 0.35, the episode cannot succeed.

Table 10: Stability deltas per mutation. Negative values indicate destabilization. Position 2 (G41) mutations are the most destabilizing but are required for high combinatorial fitness.

Pos 0 (V39)		**Pos 1 (D40)**		**Pos 2 (G41)**		**Pos 3 (V54)**	
AA	Δs	AA	Δs	AA	Δs	AA	Δs
V	0.00	D	0.00	G	0.00	V	0.00
L	-0.02	A	$+0.02$	A	-0.18	A	$+0.03$
I	-0.03	H	-0.05	C	-0.20	G	$+0.05$
F	-0.08	Y	-0.08	F	-0.22	C	-0.06
W	-0.15	W	-0.10	L	-0.25	F	-0.10

Starting sequence and observation. For seed-dependent variation, we apply one random mutation to the wild-type sequence at reset. The agent observes the current sequence, a noisy additive fitness estimate (± 0.20 Gaussian) together with a warning that it may not reflect true combinatorial fitness, a qualitative stability label, available mutations with local estimates, and recent edit history.

Finalize protocol. The agent may attempt `FINALIZE` at any time provided stability ≥ 0.35. True combinatorial fitness is revealed only upon finalization. Success requires both fitness ≥ 5.0 and stability ≥ 0.35.

C.6 Calibration Procedure and Operating-Point Selection

Each environment was calibrated to ensure that random and myopic baselines achieve low but non-zero success rates, privileged diagnostic baselines achieve high success rates, and LLM agents occupy an intermediate tier with room for prompt-induced variation.

Word Ladder. The main calibration lever is the shortest-path range $[3, 6]$. Pairs with $k_{\text{opt}} \leq 2$ are too easy, while pairs with $k_{\text{opt}} \geq 7$ yield near-zero success for all non-privileged agents.

Alloy. The key parameters are the UTS target, density target, and heavy-element scaling threshold. These were chosen so that the observation-level greedy heuristic remains above random but far below the privileged balanced policy.

GB1. We swept the fitness target across $\{4.7, 4.8, 5.0\}$ with four policies (Table 35). At all thresholds, the additive greedy baseline achieves 0%, while the privileged policy remains substantially above zero. We selected fitness_target $= 5.0$ as the main setting because it most clearly separates additive baselines, LLM agents, and privileged diagnostic baselines.

C.7 Instance Statistics

C.8 Filtering and Quality Control

Word Ladder. From 500 candidate pairs, we removed 91 pairs with shortest-path distance outside $[3, 6]$. We verified that all retained pairs are connected in the Hamming-1 graph and that neither start nor target word has zero valid neighbors at the initial state.

Alloy. The composition space is continuous, so there is no fixed instance pool. Instead, we verified that jittered starting compositions satisfy element bounds, that at least one feasible composition exists within 8 steps of each starting point, and that the surrogate model predictions are monotonic in expected directions for major elements.

GB1. The 625 variants are exhaustively enumerated from 4 positions $\times$ 5 residues. We verified that the fitness lookup table contains all 625 entries, that wild-type fitness is normalized to 1.0, and that single-site marginal estimates are computed correctly. No variants were excluded from the evaluation pool.

Table 11: Instance-level statistics for each environment. The reported quantities summarize search-space size, budget, observation noise, and the relative difficulty of random, myopic, privileged, and LLM policies. The intended calibration regime is visible in the separation between these tiers. Word Ladder and alloy-like editing admit useful but limited myopic heuristics; GB1 instead makes additive myopic reasoning fail, while privileged epistasis-aware policies remain high.

	Word Ladder	**Alloy-like**	**GB1 landscape**
Instance pool size	409 pairs	continuous	625 variants
Budget (steps)	$5k_{\text{opt}}$ (15–30)	8	6
Avg. branching factor	9.9	∼30/step	∼4/step
Observation noise	none	±30 MPa	±0.20
Random baseline success	0%	4%	4%
Observation-level heuristic success	82%	22%	0%
Best privileged-policy success	100%	93%	100%
LLM zero-shot success	18%	12%	24%
LLM best-prompt success	30%	16%	38%

C.9 Reproducibility Notes

All environments are fully deterministic given a seed. The seed controls instance selection, observation noise, and any stochastic baseline behavior. We used seeds 0–49 for the main experiments and seeds 0–99 for baseline evaluations. The original controlled LLM sweeps queried the Together hosted-model API using `Qwen/Qwen2.5-7B-Instruct-Turbo`; cross-model sanity checks queried OpenRouter-hosted instruction models with temperature 0.0 and identical environment code. The artifact directory includes raw per-episode JSON logs, summary JSON files, environment code, prompt templates, word-pair lists, alloy data, and GB1 lookup artifacts, so all reported aggregate tables can be recomputed from released files rather than from private intermediate state. No model training is performed. Hosted experiments are inference-only and are costed by episode counts, model-call budgets, and released token summaries; the non-hosted verifier and derived-table regeneration run from stored JSON logs and do not require GPUs or external API calls. Repository code is released under an MIT license; source datasets retain their original terms. The GB1 artifact is a restricted sequence-landscape proxy and contains no wet-lab protocol, synthesis instruction, DNA-design workflow, or biosynthetic design guidance. The supplementary verifier checks these aggregate links and also verifies the deterministic Alloy, GB1, and document-plan bridge diagnostics. The document-plan diagnostic is included to make an NLP-native bridge task auditable. OpenAI and open-family bridge runs are reported below as supplementary evidence, but they are not a replacement for the main three-domain recoverability measurement protocol.

C.10 Trajectory Diagnostic Definitions

For every non-final accepted edit $e_t : s_t \to s_{t+1}$ with remaining post-edit budget $b_t - 1$, the artifact records an environment-specific future-feasibility probe $R(s, b)$. Higher R always means that the state is easier to repair within the remaining budget. Word Ladder uses normalized shortest-path slack to the target; alloy-like editing uses a bounded composition-edit beam and the best attainable proxy score; GB1 uses a bounded mutation probe scored by the epistatic lookup landscape. An episode has a recoverability drop if any accepted edit satisfies

$$R(s_{t+1}, b_t - 1) < R(s_t, b_t) \tag{10}$$

or changes the probe from recoverable to unrecoverable. An episode has local deception if an accepted edit improves the task's visible local proxy while also satisfying the strict recoverability-drop inequality above. These are episode-level indicators: a trajectory contributes at most once to each diagnostic rate, regardless of how many offending edits it contains.

C.11 NLP-Native Document-Plan Bridge Experiments

To check that the constructive-editing framing is not limited to Word Ladder and scientific proxies, we added a small document-plan editing task in which a model edits a five-section outline under local struc-

tural constraints and a global target-plan constraint. This task is not included in the main three-domain aggregate tables because it is a bridge probe rather than a fourth benchmark domain.

Table 12: OpenAI document-plan bridge pilots. Temperature 0.0, seeds 0–49. Surface counts are malformed or illegal edit events; structural counts are budget, oscillation, premature-finalize, or global feasibility events.

Model	Prompt	Success	Local valid	Surface	Structural	Calls/episode
GPT-4o-mini	Zero-shot	44%	87%	28	159	5.5
GPT-4o-mini	Scaffold	72%	100%	1	133	5.3
GPT-4o	Zero-shot	98%	97%	5	87	4.7
GPT-4o	Scaffold	94%	99%	1	110	4.8

On GPT-4o-mini, the scaffold improves success by 28 percentage points and removes most surface-level failures. On GPT-4o, success is near saturation under both prompts, so the scaffold mainly reduces surface events rather than raising success. The bridge evidence is therefore no longer single-model; the open-family table below tests whether the same task is informative for smaller open-family models.

Table 13: Open-family document-plan bridge package. Temperature 0.0, seeds 0–49. The mixed outcome supports the bridge interpretation but keeps the claim model-specific.

Model	Prompt	Success	Local valid	Surface	Structural	Calls/episode
Qwen3-8B	Zero-shot	32%	53%	111	99	5.7
Qwen3-8B	Scaffold	2%	44%	140	132	6.0
Ministral-8B	Zero-shot	0%	85%	43	203	6.0
Ministral-8B	Scaffold	0%	95%	13	170	6.0
Llama-8B	Zero-shot	10%	81%	50	173	6.0
Llama-8B	Scaffold	24%	81%	48	143	6.0

The open-family run is mixed rather than uniformly positive: Qwen3 succeeds most often under zero-shot prompting, Llama benefits from the scaffold, and Ministral remains at 0/50 under both prompts despite high local-valid edit rates. This closes the open-family bridge check as boundary evidence while preserving the paper's main claim on the matched Word Ladder, alloy-like, and GB1 constructive-editing proxy.

C.12 Single-Model Stress-Test Pilots

We also ran small OpenAI pilots for reviewer-raised alternatives: independent restarts, shallow search wrappers, surrogate sensitivity, memory controls, and date-separated hosted-model API stability. These pilots are not used as main evidence because they are small-sample and mostly single-model. They are included to make the remaining empirical risks concrete and auditable.

Table 14: Single-model search-wrapper pilots. Independent restarts use three attempts per task seed. Shallow wrappers use `text_visible` scoring, $k = 3$, and beam depth 1.

Pilot	Environment	Success	Calls/episode	Structural
Restarts	Alloy	1/10	24.0	265
Restarts	GB1	0/10	15.7	72
Greedy	Word Ladder	3/5	18.2	5
Self-consistency	Word Ladder	3/5	43.8	2
Beam	Word Ladder	1/5	55.2	10
Loop-avoidant	Word Ladder	2/5	93.0	3
Greedy	Alloy	0/5	8.0	44
Self-consistency	Alloy	1/5	24.0	33
Beam	Alloy	1/5	24.0	40
Loop-avoidant	Alloy	0/5	40.0	32
Greedy	GB1	0/5	5.0	9
Self-consistency	GB1	0/5	15.0	9
Beam	GB1	0/5	13.2	8
Loop-avoidant	GB1	0/5	25.0	9

The GPT-4o-mini wrapper pilots show that extra calls can improve local validity or help occasionally on Alloy, but they do not consistently raise success: Word Ladder self-consistency matches greedy success while reducing structural events, beam underperforms greedy on Word Ladder, loop avoidance remains below greedy on Word Ladder and at 0/5 on Alloy/GB1, and GB1 remains unsolved in this small sample. A targeted GPT-4o wrapper pilot over the same three domains, seeds 0–4, `text_visible` scoring, $k = 3$, and beam depth 1 gives a stronger-model contrast: Word Ladder reaches 2/5 greedy, 3/5 self-consistency, 1/5 beam, and 3/5 loop-avoidant; Alloy reaches 2/5, 4/5, 3/5, and 4/5; GB1 remains 0/5 under all four controllers despite 100% local validity. Thus shallow wrappers can materially help a stronger model on the controlled compositional proxy, but do not remove the GB1 landscape failure in this pilot.

We also ran an open-family shallow-wrapper checkpoint over Qwen3-8B, Ministral-8B, and Llama-3.1-8B across the three main environments using the same visible-state scorer and measured call/token accounting. Table 15 summarizes the main controller-level takeaways; full controller-aware rows and raw-log pointers are in the supplement. The clean checkpoint excludes three quarantined network/API-error logs, so Qwen Word Ladder beam has 49 clean seeds and Qwen Alloy greedy has 48 clean seeds. Overall, shallow wrappers are not a universal repair: they help Llama Alloy substantially and sometimes improve Word Ladder, but GB1 remains governed by local-optimum traps even when local validity is near perfect.

Table 15: Open-family shallow-wrapper checkpoint. Values are success rates over 50 seeds unless marked clean-N after quarantining network/API-error raw logs. Best wrapper excludes deterministic greedy; calls/episode are shown for the best wrapper.

Model	Environment	Greedy success	Best wrapper success	Calls
Qwen3-8B	Word Ladder	44%	49% beam, $N = 49$	153.8
Qwen3-8B	Alloy	10%, $N = 48$	12% self/loop	39.9
Qwen3-8B	GB1	12%	12% self	26.2
Ministral-8B	Word Ladder	56%	62% self/beam	69.7
Ministral-8B	Alloy	4%	20% self	40.0
Ministral-8B	GB1	28%	24% beam	61.9
Llama-8B	Word Ladder	46%	60% self/loop	68.6
Llama-8B	Alloy	6%	34% self	38.8
Llama-8B	GB1	22%	26% loop	27.1

The same checkpoint reinforces the cost-sensitivity of any wrapper claim. For example, Llama Alloy self-consistency reaches 17/50 successes versus 3/50 for greedy, but uses 38.8 calls/episode instead of 8.0. Llama GB1 loop avoidance is the best Llama GB1 row at 13/50 successes, but still leaves LOCAL-OPTIMUM-TRAP as the dominant terminal failure and costs about 4.8 times greedy's calls. Thus the revised claim is not that search wrappers fail, but that shallow visible-state wrappers only partially close the structural gap and do so unevenly across domains.

We then ran a costlier depth-3 visible-state beam stress test for the same three open-family small models. This is still not an optimized planner: it samples candidate actions from the model, scores visible next states, and expands a depth-limited tree without exact recoverability information, rollback, or a learned value function. Table 16 shows that the extra search is not a monotone repair. It raises some Word Ladder and Alloy cells, but leaves GB1 at 8–18% success with LOCAL-OPTIMUM-TRAP still dominant. The result supports the narrower interpretation used in the main text: GB1 requires the right trajectory-level information or backtracking policy, not merely more sampled visible-state lookahead.

To separate more sampled LLM continuations from explicit search architecture, we also ran a non-LLM MCTS-style boundary check. The planner enumerates valid edits from the simulator and uses 64 simulations per decision. In visible mode, rollouts are valued by the same visible proxy family used in the task observation; in privileged mode, rollouts use terminal-objective information and should be read only as a diagnostic upper bound. Table 17 shows two useful boundaries. First, visible-proxy MCTS solves many GB1 seeds but still leaves 13/50 local-optimum failures, confirming that GB1 is not just an action-format problem. Second, privileged MCTS solves GB1 completely and substantially improves Alloy, showing that explicit state search with the right objective information can change the intervention profile. This is

Table 16: Open-family depth-3 visible-state beam stress test. All rows use the structural prompt, state-only memory, `text_visible` scoring, $k = 3$, beam width 4, beam depth 3, and 50 seeds. Calls are average model calls per episode. Surface-clean success is conditioned on episodes without surface failures.

Model	Environment	Success	Surface-clean success	Calls
Qwen3-8B	Word Ladder	20/50	11/15	156.3
Qwen3-8B	Alloy	5/50	4/45	80.0
Qwen3-8B	GB1	5/50	5/47	40.4
Ministral-8B	Word Ladder	24/50	12/13	263.6
Ministral-8B	Alloy	7/50	7/40	122.6
Ministral-8B	GB1	4/50	4/50	52.4
Llama-8B	Word Ladder	26/50	11/12	281.8
Llama-8B	Alloy	10/50	9/48	151.0
Llama-8B	GB1	9/50	9/49	85.0

why we do not interpret LLM wrapper failures as evidence that no planner could solve the tasks.

Table 17: Non-LLM MCTS-style boundary check over seeds 0–49. Visible mode uses visible-proxy rollout value; privileged mode uses terminal-objective information and is an upper bound, not an attainable LLM policy. Each decision uses 64 MCTS simulations.

Environment	Mode	Success	Mean steps	Dominant residual failure
Alloy	Visible proxy	3/50	7.58	hard constraint
Alloy	Privileged	41/50	6.04	hard constraint
GB1	Visible proxy	37/50	3.72	local optimum
GB1	Privileged	50/50	2.38	–

Separately, we ran a small reasoning-model boundary pilot on GB1 with o4-mini through the same OpenRouter chat-completions interface. Earlier low-token smoke tests produced empty actions, so the reported row uses `max_tokens`=2048. The pilot remains small and includes four malformed actions, so we do not treat it as model-family evidence. It is nevertheless a useful boundary check: among parsed edit attempts, local validity is 23/23, yet all five episodes fail with LOCAL-OPTIMUM-TRAP. This is consistent with the GB1 mechanism claim but is intentionally kept out of the main results.

Table 18: Small o4-mini GB1 reasoning-model boundary pilot. The run uses the structural prompt, state-only memory, greedy actions, seeds 0–4, and `max_tokens`=2048. Because the run still has malformed actions, it is a compatibility and boundary check rather than main evidence.

Model	Success	Local-valid edits	Surface-clean episodes	Terminal failure	Calls
o4-mini	0/5	23/23	2/5	5/5 local-optimum trap	5.6

The sensitivity sweep shows setting dependence but not task collapse: relaxing Alloy heavy-scaling and using the GB1 additive evaluator both raise success, yet absolute success remains low and structural events remain common. The memory sweep suggests that history perturbations can change behavior, especially on Word Ladder, but full history is not a monotone fix and GB1 remains hard under all memory conditions. Summary and best-state memory are also not clean fixes: summary reaches 4/20 on Alloy but best-state drops to 1/20, and neither changes the GB1 conclusion. The Alloy gains under randomized and misleading history should be interpreted as variance or prompt perturbation evidence, not as evidence that informative memory solves the task. In a targeted GPT-4o sensitivity stress test over seeds 0–9, Alloy reaches 5/10 under both the default and heavy-scaling-off settings, while GB1 remains 0/10 under both the default and additive-evaluator settings. Thus the stronger model reduces the Alloy setting dependence visible in GPT-4o-mini but does not remove the GB1 failure in this small sample. A targeted GPT-4o memory stress test over seeds 0–4 similarly keeps the memory conclusion cautious: Word Ladder rises from 2/5 state-only to 5/5 full-history, but also reaches 4/5 under randomized history and 3/5 under misleading history; Alloy moves from 2/5 state-only to 1/5 full-history and 3/5 under randomized or misleading history; GB1 remains 0/5 under all four memory conditions. Extra context can therefore

Table 19: Single-model sensitivity and memory-control pilots. Sensitivity rows use GPT-4o-mini, the structural prompt, and seeds 0–49; memory-control rows use GPT-4o-mini, the structural prompt, and seeds 0–19.

Pilot	Condition	Success	Local valid	Structural
Alloy sensitivity	Default	4/50	77%	415
Alloy sensitivity	Heavy scaling off	7/50	76%	384
GB1 sensitivity	Default	4/50	100%	115
GB1 sensitivity	Additive evaluator	9/50	100%	59
Word Ladder memory	State only	6/20	32%	21
Word Ladder memory	Full history	9/20	70%	13
Word Ladder memory	Summary	7/20	46%	14
Word Ladder memory	Best-state	8/20	52%	13
Word Ladder memory	Randomized history	11/20	70%	10
Word Ladder memory	Misleading history	11/20	74%	11
Alloy memory	State only	0/20	78%	188
Alloy memory	Full history	2/20	72%	162
Alloy memory	Summary	4/20	71%	154
Alloy memory	Best-state	1/20	73%	166
Alloy memory	Randomized history	2/20	74%	162
Alloy memory	Misleading history	3/20	69%	156
GB1 memory	State only	2/20	100%	44
GB1 memory	Full history	0/20	100%	46
GB1 memory	Summary	2/20	100%	46
GB1 memory	Best-state	2/20	100%	56
GB1 memory	Randomized history	0/20	100%	52
GB1 memory	Misleading history	1/20	100%	51

perturb or improve behavior, but it is not a reliable substitute for trajectory-level search over the latent landscape.

Table 20: Date-separated OpenAI document-plan stability pilot on matched seeds 0–19. The 2026-05-10 rows are a subset of the 50-seed bridge pilot; the 2026-05-11 rows are a fresh rerun.

Date	Prompt	Success	Local valid	Structural
2026-05-10	Zero-shot	10/20	84%	58
2026-05-10	Scaffold	11/20	100%	59
2026-05-11	Zero-shot	13/20	88%	56
2026-05-11	Scaffold	14/20	99%	61

The date-separated rerun preserves the qualitative direction of the scaffold effect but shifts both success rates by 15 percentage points. This is why we avoid fine-grained hosted-model API ranking claims and release raw logs and run dates for auditability.

Table 21: Same-day OpenAI GPT-4o-mini main-domain API stability rerun on matched seeds 0–19. Both runs use deterministic decoding.

Environment	Prompt	Pass A	Pass B	Delta
Word Ladder	Zero-shot	40%	30%	-10
Word Ladder	Scaffold	40%	30%	-10
Alloy	Zero-shot	25%	0%	-25
Alloy	Self-check	5%	10%	+5
GB1	Zero-shot	5%	5%	0
GB1	Self-check	5%	5%	0

A same-day rerun on the three main domains shows larger absolute movement for Word Ladder and Alloy than for GB1. The qualitative GB1 pattern is stable: local edit validity remains 100%, but success remains near zero. These results reinforce the reporting choice to emphasize qualitative patterns, paired seeds, and raw logs rather than fine-grained hosted-model API rankings.

This sampled-decoding pilot remains single-model and should not by itself be treated as open-family robustness evidence. It does show why deterministic decoding is a protocol choice: sampled decoding

Table 22: OpenAI GPT-4o-mini decoding-robustness pilot on the three main domains, matched seeds 0–19. Deterministic uses temperature 0.0; sampled uses temperature 0.7 and top-p 0.9.

Decoding	Environment	Prompt	Success	Local valid	Structural
Deterministic	Word Ladder	Zero-shot	8/20	42%	14
Sampled	Word Ladder	Zero-shot	7/20	33%	13
Deterministic	Word Ladder	Scaffold	8/20	39%	13
Sampled	Word Ladder	Scaffold	7/20	39%	20
Deterministic	Alloy	Zero-shot	5/20	66%	159
Sampled	Alloy	Zero-shot	2/20	65%	178
Deterministic	Alloy	Self-check	1/20	79%	174
Sampled	Alloy	Self-check	0/20	70%	168
Deterministic	GB1	Zero-shot	1/20	100%	47
Sampled	GB1	Zero-shot	0/20	100%	46
Deterministic	GB1	Self-check	1/20	100%	45
Sampled	GB1	Self-check	1/20	100%	49

leaves Word Ladder close in success but lowers zero-shot local validity, reduces Alloy success, and leaves GB1 at 100% local edit validity but near-zero success under both decoding settings. The open-family closure panel below adds the completed sampled decoding robustness grid.

C.13 Open-Family Diagnostic Closure Panel

We complement the single-model pilots above with a five-experiment open-family panel that uses the same three small-block open-weight models (Qwen3-8B, Ministral-8B, Llama-3.1-8B) across the three main environments. All runs use deterministic decoding unless noted, the structural prompt where one exists, and either 20 or 50 seeds depending on the budget guard documented in the supplement. The panel is intended to convert the pilot-scale claims above into open-family evidence rather than to overturn them. The aggregator script is at `supplementary/code/aggregate_award_closure.py`; raw episode logs live under `results_v4/`.

Table 23: Environment-specific best model on the 13-model open-family panel (structural prompt, 50 seeds, deterministic decoding). Three different families win the three environments; sizes differ as well.

Environment	Best model	Family	Success
Word Ladder	Llama-3.1-8B	Llama	46%
Alloy	Llama-3.3-70B	Llama	26%
GB1	Ministral-8B	Mistral	22%

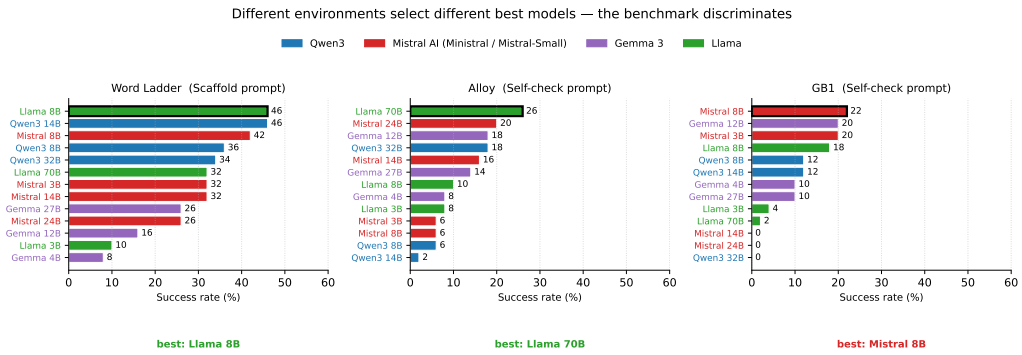

Figure 3: Different environments select different best models in the 13-model open-family panel under the structural prompt. The panel is diagnostic scope evidence, not a leaderboard or scaling-law claim.

Scorer ablation. Figure 4 and Table 25 report the same shallow-wrapper grid as Table 15 but varies the scorer between `text_visible`, `proxy`, and `oracle`. The first two scorers use visible-state information; `oracle` is a privileged diagnostic sensitivity condition, not an attainable agent policy. The privileged

Table 24: Three benchmark properties supported by the open-family closure panel. *Discrimination* shows that the benchmark separates models; *stability* shows that protocol choices do not flip the verdict; *sensitivity* shows that interventions reveal model-specific behavior.

Property	Evidence (open-family, 13 or 3 models)	Pointer
Discrimination	Best model differs across the three environments	Tab. 23
	Within-family best size differs (Qwen3 14B > 32B on Word Ladder)	Fig. 7
	Llama 70B leads Alloy but is last on GB1	Fig. 8
Stability (no qualitative flip)	Scorer ablation: cell-level Δ success up to $\approx$25pp, no qualitative regime flip	Tab. 25, Fig. 4
	Sampled vs. deterministic decoding: GB1 stays at $\approx$100% local validity	Tab. 29
	Hosted-model API rerun: no environment changes regime	Tab. 28
Sensitivity (model-specific)	Best memory regime differs across all 9 (model, env) cells	Tab. 26, Fig. 5
	Surrogate toggles shift levels but preserve low-success regime	Tab. 27, Fig. 6
	Prompt-by-environment interaction (cross-model)	Tab. 31

condition is not monotone: Llama Word Ladder drops from 60% under `text_visible` to 45% under `oracle`. This supports the reporting choice to treat scorer choice as a sensitivity analysis rather than as a headline performance claim. Llama Alloy is the only cell where privileged scoring clearly outperforms visible scoring (40% versus 15%), suggesting that distance-aware scoring closes a measurable gap mainly in the alloy-like proxy.

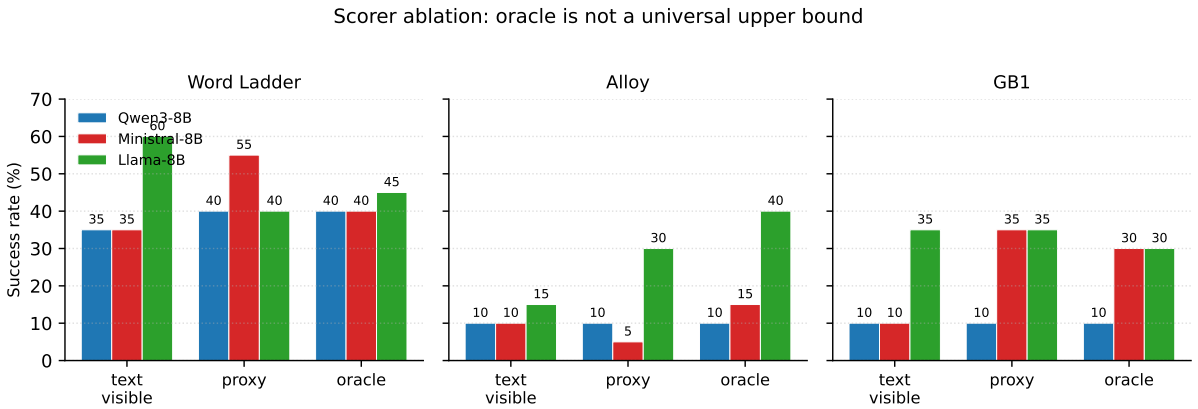

Figure 4: Open-family scorer ablation. Each panel shows three scorers (text_visible, proxy, privileged) over 20 seeds for the three small-block models. Bars are not monotone left-to-right: Llama Word Ladder drops from 60% under text_visible to 45% under the privileged scorer, demonstrating that privileged target-distance scoring is a sensitivity condition rather than an attainable monotone upper bound.

Extended memory grid. Figure 5 and Table 26 expand the memory-control pilot to all three environments at 50 seeds per cell. Several patterns are stable. Word Ladder rewards longer or shuffled history for Ministral (62% randomized, 60% full) while Llama prefers the bounded `Window-3` regime (54%) and is hurt by full history (30%). Alloy yields the smallest absolute gains and the biggest within-model swings: Llama jumps from 8% under state-only to 22% under randomized history, whereas Ministral's full-history advantage of 14% is matched by misleading history at only 6%. GB1 stays in the low-success regime under every memory condition (best cell 32%, Llama best-state), confirming that the landscape failure is not a memory problem. Across environments, no single memory module dominates: the best memory depends on both model and environment, which is why the main paper treats `state_only` as the conservative reporting choice.

Surrogate sensitivity. Figure 6 and Table 27 report the open-family version of the GPT-4o-mini sensitivity sweep. The Alloy heavy-scaling toggle raises mean success modestly (Qwen3 zero-shot from

Table 25: Open-family scorer ablation for shallow search wrappers (beam, $k = 3$, depth 1). Each cell is success rate over 20 seeds; `text_visible` and `proxy` use only information already exposed to the policy, while `oracle` uses privileged target-distance information and is reported only as a sensitivity condition. Calls/episode in the rightmost column are for the `text_visible` scorer.

Model	Environment	Text-visible	Proxy	Privileged	Calls/episode
Qwen3-8B	Word Ladder	35%	40%	40%	50.7
Qwen3-8B	Alloy	10%	10%	10%	23.9
Qwen3-8B	GB1	10%	10%	10%	16.4
Ministral-8B	Word Ladder	35%	55%	40%	52.2
Ministral-8B	Alloy	10%	5%	15%	24.0
Ministral-8B	GB1	10%	35%	30%	16.8
Llama-8B	Word Ladder	60%	40%	45%	44.0
Llama-8B	Alloy	15%	30%	40%	24.0
Llama-8B	GB1	35%	35%	30%	16.5

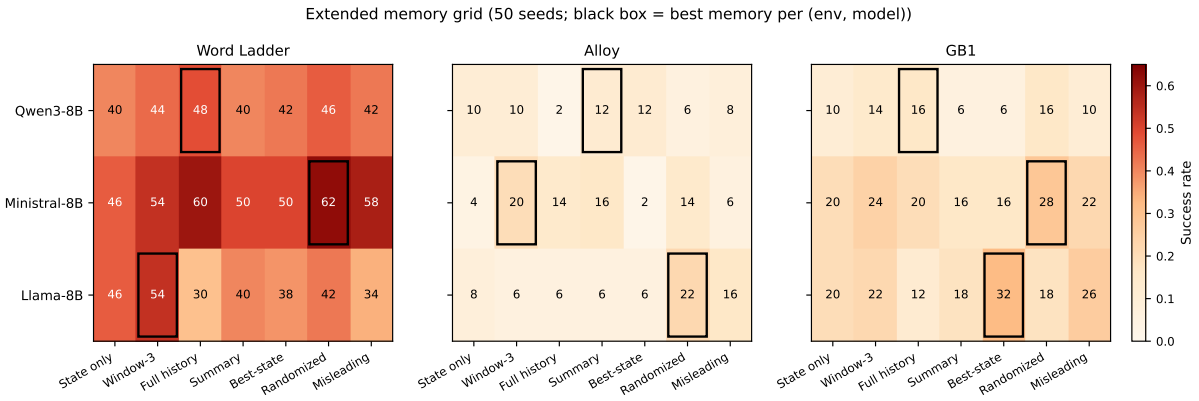

Figure 5: Open-family extended-memory grid (50 seeds, structural prompt, deterministic decoding). Black boxes mark the best memory module per (env, model). The best memory differs across every row in every panel, so no single memory module dominates: `Full history` is best for Qwen3 on Word Ladder, `Window-3` for Llama, and `Randomized` for Ministral; the alignment changes again on Alloy and GB1.

Table 26: Open-family extended-memory grid. Each cell is success rate over 50 seeds, structural prompt, deterministic decoding. `Randomized` shuffles the trajectory history before each call; `Misleading` injects a length-matched but reward-mismatched history. The state-only column matches the main paper's deterministic baseline.

Environment	Model	State only	Window-3	Full history	Summary	Best-state	Randomized	Misleading
Word Ladder	Qwen3-8B	40%	44%	48%	40%	42%	46%	42%
	Ministral-8B	46%	54%	60%	50%	50%	62%	58%
	Llama-8B	46%	54%	30%	40%	38%	42%	34%
Alloy	Qwen3-8B	10%	10%	2%	12%	12%	6%	8%
	Ministral-8B	4%	20%	14%	16%	2%	14%	6%
	Llama-8B	8%	6%	6%	6%	6%	22%	16%
GB1	Qwen3-8B	10%	14%	16%	6%	6%	16%	10%
	Ministral-8B	20%	24%	20%	16%	16%	28%	22%
	Llama-8B	20%	22%	12%	18%	32%	18%	26%

2% to 12%, Llama zero-shot from 0% to 10%) but does not change the qualitative low-success pattern. The GB1 additive evaluator does the opposite of the pilot: it lowers success for every model (Ministral self-check drops from 24% to 2%; Llama from 22% to 0%), which shows that the GB1 difficulty in the main results is not an artifact of the multiplicative evaluator. Both directions are inside the pilot's claim that surrogate settings can perturb behavior but do not collapse the qualitative structure.

Hosted-model API stability. Table 28 extends the date-separated API stability check from the document-plan pilot (Table 20) to the main-domain open-family panel. Cell-level absolute success varies

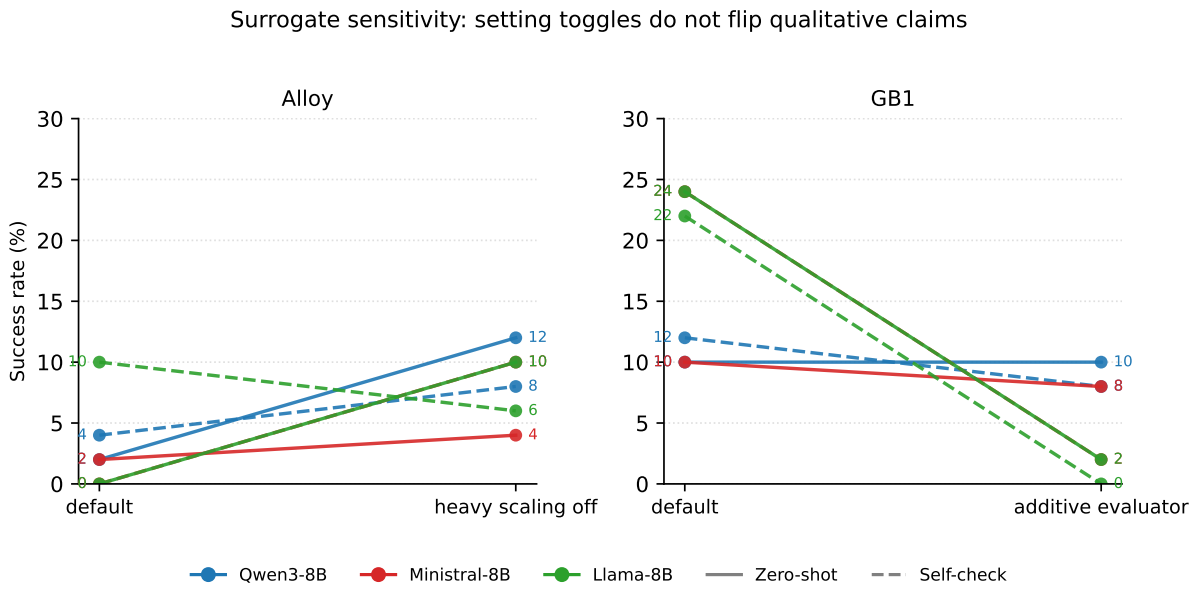

Figure 6: Open-family surrogate-sensitivity slope chart. Lines connect each (model, prompt) cell's success rate under the default and perturbed settings. The Alloy heavy-scaling toggle lifts most cells; the GB1 additive evaluator drops most cells. Both directions stay inside the low-success regime, which is the property the main paper relies on.

Table 27: Open-family surrogate sensitivity on Alloy (heavy-scaling toggle) and GB1 (additive evaluator). Each cell is success rate over 50 seeds; ZS is zero-shot, SC is self-check.

Environment	Setting	Model	Zero-shot	Self-check
Alloy	default	Qwen3-8B	2%	4%
		Ministral-8B	2%	0%
		Llama-8B	0%	10%
	heavy scaling off	Qwen3-8B	12%	8%
		Ministral-8B	4%	10%
		Llama-8B	10%	6%
GB1	default	Qwen3-8B	10%	12%
		Ministral-8B	10%	24%
		Llama-8B	24%	22%
	additive evaluator	Qwen3-8B	10%	8%
		Ministral-8B	8%	2%
		Llama-8B	2%	0%

by up to 25 percentage points relative to other deterministic Alloy and GB1 runs in this appendix, which is why the main paper avoids fine-grained model rankings on hosted APIs and reports qualitative patterns and raw logs instead.

Table 28: Open-family hosted-model API stability rerun on Alloy and GB1 (20 seeds, deterministic decoding, 2026-05-15). Each cell is success rate; ZS is zero-shot, SC is self-check.

Environment	Model	Zero-shot	Self-check
Alloy	Qwen3-8B	0%	15%
	Ministral-8B	10%	0%
	Llama-8B	5%	0%
GB1	Qwen3-8B	10%	10%
	Ministral-8B	15%	40%
	Llama-8B	30%	15%

Decoding robustness. Table 29 reports the open-family sampled-decoding panel (temperature 0.7, top-p 0.9, 20 seeds). Compared against the deterministic main-results panel for the same models and prompts, sampled decoding does not flip any of the central qualitative claims: Word Ladder remains the highest-success environment, Alloy and GB1 remain in the low-success regime, and GB1 keeps approximately 100% local edit validity under both decoding modes. Within-cell absolute success moves by up to roughly 20 percentage points for the smaller models, which is again why the main paper emphasizes paired-seed reporting and matched protocol comparisons rather than single-cell ranks.

Table 29: Open-family decoding-robustness panel. Sampled decoding uses temperature 0.7 and top-p 0.9; 20 seeds per cell. Compare against the deterministic main-results panel for the same model and prompt.

Environment	Model	Zero-shot	Self-check/scaffold
Word Ladder	Qwen3-8B	45%	40%
	Ministral-8B	15%	30%
	Llama-8B	35%	35%
Alloy	Qwen3-8B	0%	5%
	Ministral-8B	0%	5%
	Llama-8B	10%	10%
GB1	Qwen3-8B	10%	10%
	Ministral-8B	0%	10%
	Llama-8B	15%	15%

D Detailed Main Prompt Results

This section provides the full prompt-intervention table underlying the main comparison in Sections 5.5 and 5.3. The main pattern to verify is that format-only few-shot prompting does not consistently improve performance, while larger gains appear only when the intervention changes the editing policy in an environment-specific way.

Table 30: Detailed main prompt-intervention results under state-only history ($M0$).

Environment	Zero-shot	Few-shot (format)	Few-shot (strategy)	Scaffold	Self-check (generic)	Self-check (task-specific)
Word Ladder (4-letter)	18 [10,31]	8 [3,19]	18 [10,31]	30 [19,44]	6 [2,16]	18 [10,31]
Alloy-like	12 [6,24]	10 [4,21]	2 [0,10]	14 [7,26]	6 [2,16]	16 [8,29]
GB1 landscape	24 [14,37]	20 [11,33]	18 [10,31]	12 [6,24]	26 [16,40]	38 [26,52]

E Cross-Model Sanity Check

This appendix reports the cross-model sanity check used to scope the main claims beyond a single model. We evaluate six additional instruction models under $M0$ on the three environments. For each environment, we compare zero-shot, format-only few-shot, and the environment-specific structural prompt (scaffold for Word Ladder; task-specific self-check for alloy and GB1). Each cell uses 50 seeds.

Aggregating across these 54 cells gives 2,700 additional episodes. In the two scientific-design-inspired proxies, structural events dominate surface events by a wide margin: 7,276 versus 187 in alloy, and 1,866 versus 321 in GB1. This supports the diagnostic claim that the benchmark exposes failures in delayed trajectory construction, not only action formatting. The cross-model table also shows why the claim must be scoped carefully: the best prompt is not stable across all model families, and larger models do not uniformly dominate smaller ones in these constrained editing environments.

The artifact also includes expanded family/scale summaries beyond this compact appendix table: a Qwen/Mistral/Gemma tiny-to-large synthesis and a separate Llama 3B/8B/70B scaling block, each with raw logs, bootstrap intervals, surface-conditioned success, and trajectory diagnostics. We keep those analyses out of the main table because the model grid is intentionally unbalanced across families; they support qualitative recurrence, not a scaling law or leaderboard.

Table 31: Cross-model success rates (%) under zero-shot / format-only / structural prompting. Absolute performance varies across families and scales, but format-only prompting is not a consistent improvement mechanism, and alloy/GB1 failures remain dominated by delayed structural events rather than malformed actions.

Model	Word Ladder	Alloy	GB1
Gemma-3-27B-IT	28 / 30 / 26	22 / 12 / 14	4 / 4 / 10
Llama-3.1-8B-Instruct	48 / 54 / 46	6 / 8 / 10	24 / 22 / 18
Llama-3.2-3B-Instruct	10 / 12 / 10	24 / 24 / 8	8 / 2 / 4
Llama-3.3-70B-Instruct	34 / 38 / 32	10 / 18 / 26	8 / 4 / 2
Mistral-Small-3.2-24B	34 / 36 / 26	16 / 16 / 20	0 / 6 / 0
Qwen3-32B	50 / 60 / 34	24 / 16 / 18	2 / 0 / 0

Figures 7 and 8 visualize the de-duplicated cross-family panel underlying these synthesis tables: 13 distinct models drawn from Qwen3, Mistral / Ministral, Gemma 3, and Llama, at sizes ranging from 3B to 70B parameters, evaluated on the three environments at 50 seeds per (model, env, prompt) cell. Figure 7 plots success rate against model size on a log axis for the structural prompt of each environment. Word Ladder shows visible within-family movement with size (Qwen3 peaks at 14B and Mistral at 8B, while Gemma 3 climbs monotonically), but the two scientific-design-inspired proxies stay in the low-success regime even at 27B or 70B. Figure 8 shows the same data as a model-by-environment heatmap with families color-coded; no row is uniformly best across the three columns, which is the same scope qualification stated in the Section E prose: cross-family scaling does not produce a single dominant model or a clean scaling law on the scientific-design environments.

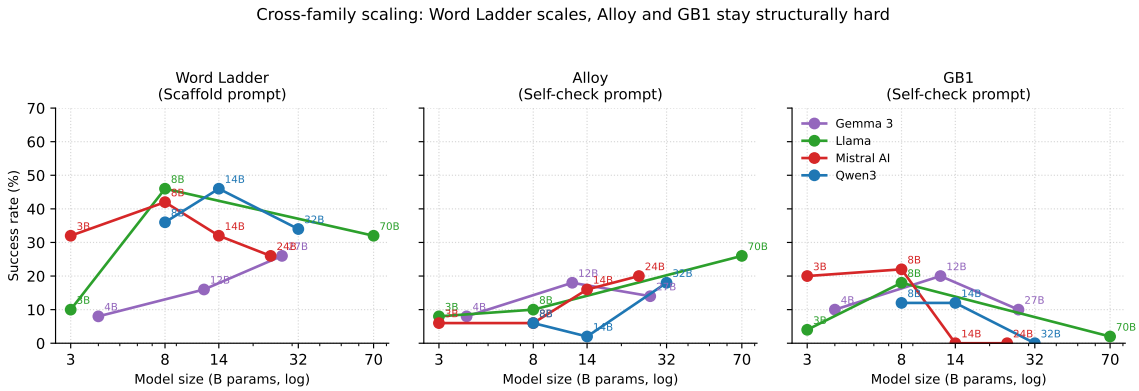

Figure 7: Cross-family scaling curves. Each panel plots success rate against model size (log scale) for one environment under the structural prompt. Line markers annotate each model's parameter count. Mistral / Ministral are merged into a single vendor line because they share the same provider but span different model series.

F Detailed History Ablation

This section reports the full history-ablation results summarized in Section 5.6. The key pattern is that recent edit history is not uniformly beneficial: it helps when the task rewards cumulative state tracking, but harms performance when recent trajectories reinforce a poor local path.

Table 32: History-ablation results for selected prompt conditions.

Environment	Prompt	$M0$	$M1$	$M3$	Interpretation
Word Ladder	zero-shot	22	20	10	harmful
Word Ladder	scaffold	30	18	10	strongly harmful
Alloy	zero-shot	12	20	25	helpful
Alloy	self-check	18	8	18	non-monotonic
GB1	zero-shot	25	8	8	strongly harmful
GB1	self-check	35	5	5	catastrophically harmful

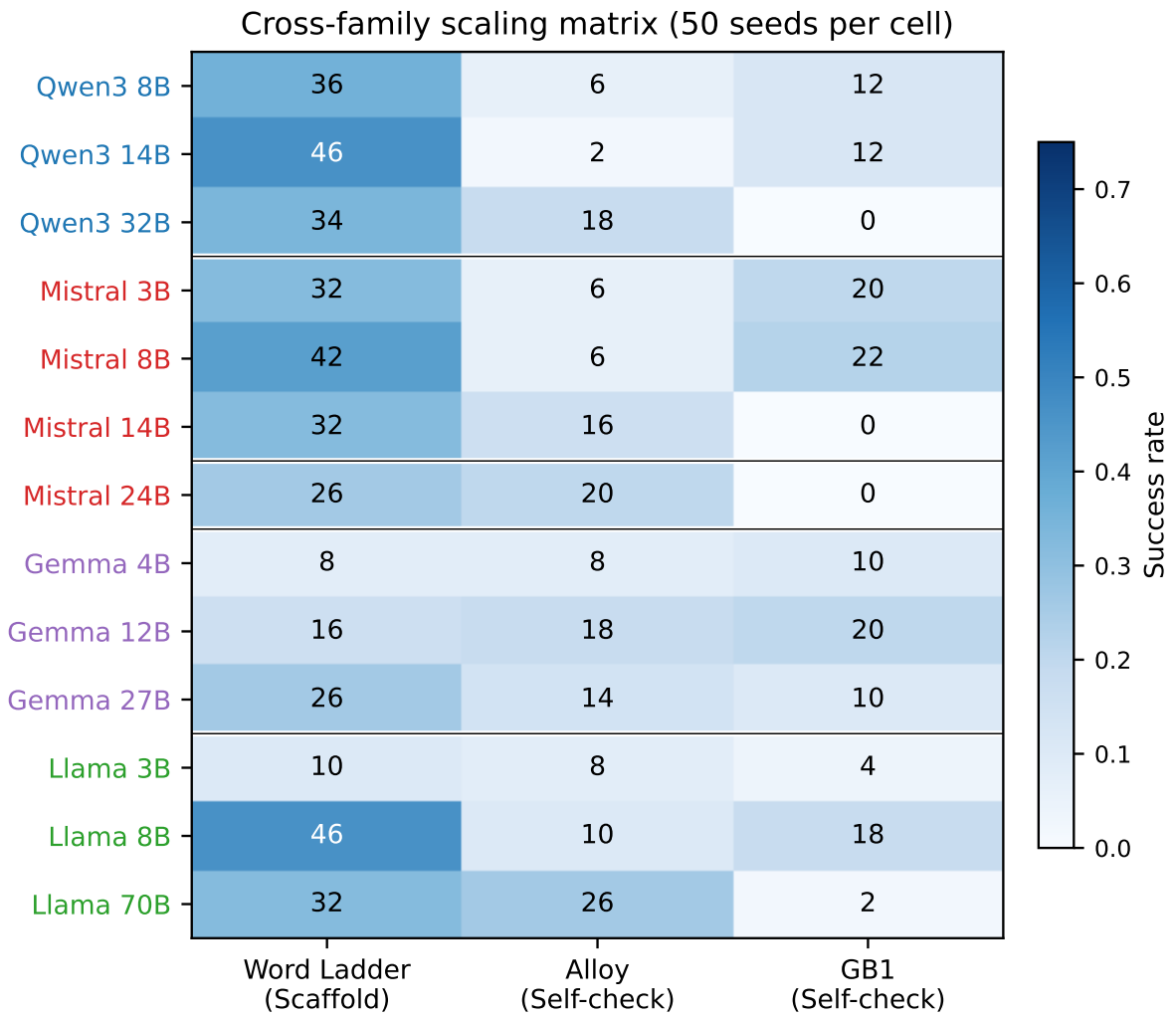

Figure 8: Cross-family scaling matrix: 13 models grouped by family and size against the three environments under the structural prompt, 50 seeds per cell. The y-axis labels are color-coded by family. No row dominates: Llama 70B leads on Alloy, Mistral 8B leads on GB1, and Qwen3 14B ties Llama 8B at the top of Word Ladder.

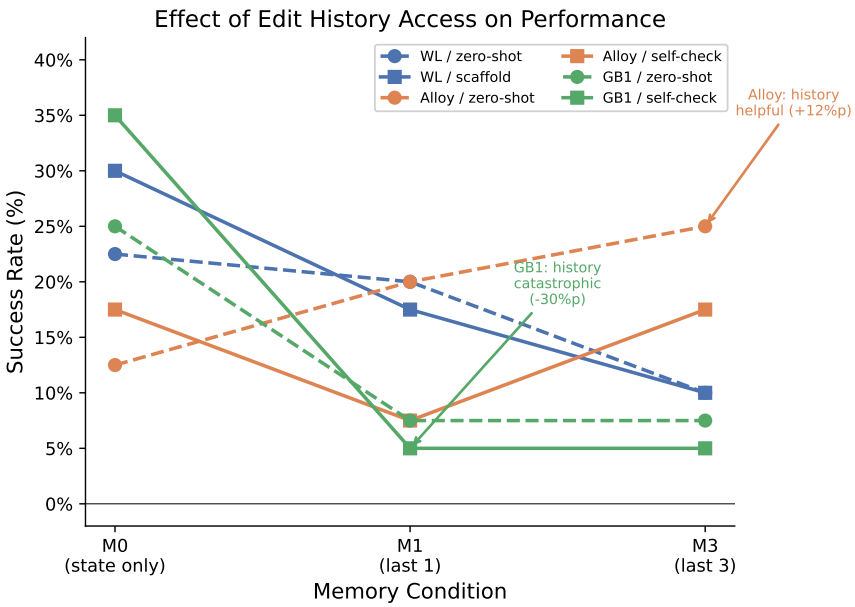

Figure 9: History ablation for selected prompt conditions. Recent trajectory access helps only when the raw history carries useful cumulative signal; it can also reinforce poor local commitments.

G Failure Patterns

This section gives the dominant failure patterns underlying the discussion in Section 5.2. The table is intended to show that the main residual errors are not uniform across environments and are predominantly structural rather than syntactic.

Table 33: Representative dominant failures under the main prompt experiment. The table shows that residual errors are predominantly structural rather than purely syntactic, and that the dominant failure mode differs by environment: budget-unaware continuation in Word Ladder, objective-tradeoff collapse in alloy-like composition editing, and local-optimum trapping in GB1. Counts in parentheses indicate the number of episodes in which the listed failure was dominant.

Environment	Prompt condition	Dominant failure
Word Ladder	zero-shot	BUDGET-UNAWARE (19)
Word Ladder	few-shot (format)	BUDGET-UNAWARE (23)
Word Ladder	few-shot (strategy)	BUDGET-UNAWARE (19)
Word Ladder	scaffold	BUDGET-UNAWARE (17)
Word Ladder	self-check (generic)	BUDGET-UNAWARE (18), PREMATURE-FINALIZE (8)
Word Ladder	self-check (task-specific)	BUDGET-UNAWARE (20)
Alloy	zero-shot	OBJECTIVE-TRADEOFF (20)
Alloy	few-shot (format)	OBJECTIVE-TRADEOFF (21)
Alloy	few-shot (strategy)	OBJECTIVE-TRADEOFF (23)
Alloy	scaffold	OBJECTIVE-TRADEOFF (22)
Alloy	self-check (generic)	OBJECTIVE-TRADEOFF (20)
Alloy	self-check (task-specific)	OBJECTIVE-TRADEOFF (18)
GB1	zero-shot	surface: 0, structural: 21; LOCAL-OPTIMUM TRAP (21)
GB1	few-shot (format)	surface: 1, structural: 25; LOCAL-OPTIMUM TRAP (24)
GB1	few-shot (strategy)	surface: 0, structural: 10; LOCAL-OPTIMUM TRAP (10)
GB1	scaffold	surface: 9, structural: 19; LOCAL-OPTIMUM TRAP (17)
GB1	self-check (generic)	surface: 0, structural: 23; LOCAL-OPTIMUM TRAP (22)
GB1	self-check (task-specific)	surface: 0, structural: 22; LOCAL-OPTIMUM TRAP (20)

H Example Prompt Templates

This section includes abbreviated prompt templates for the three environments. They are provided to make the intervention design concrete and to clarify the difference between action-format constraints and policy-shaping content.

H.1 Word Ladder as Lexical Editing

System:

```
You are solving a constructive lexical editing task.
At each step, edit exactly one character of the current word.
Your output must be exactly one action.
Use the format: EDIT <position> <letter>.
Do not explain your answer.
```

User:

```
Target: ruly
Current: many
Used path: many
Editable positions and candidate letters:
0 -> {r, w, z}
1 -> {o, u}
```

```
2 -> {l, r}
3 -> {e, s}
Remaining budget: 8
Your action:
```

H.2 Alloy as Compositional Editing

System:

```
You are solving a constructive alloy-editing task.
At each step, increase one element and decrease another by the same amount.
Use exactly one action in the format:
EDIT <inc_element> <dec_element> <delta>
or FINALIZE.
Do not explain your answer.
```

User:

```
Current composition: Fe 25, Ni 20, Cr 20, Co 15, Mn 10, Mo 10
UTS estimate: 1,850 MPa (target >= 2,000)
Density estimate: 8.35 (target <= 8.20)
Trajectory stability: moderate
Reversibility warning: low
Phase: coarse
Allowed delta: {10, 15}
Remaining budget: 6
Your action:
```

H.3 Restricted GB1 Sequence-Landscape Editing

System:

```
You are solving a constructive restricted GB1 sequence-landscape editing
task.
At each step, replace one token in the current sequence.
Use exactly one action in the format:
EDIT <position> <token>
or FINALIZE.
Do not explain your answer.
```

User:

```
Current sequence: V39=V D40=W G41=G V54=V
Estimated fitness: 3.85 (target >= 5.0)
Single-site estimates may be misleading due to epistasis.
Stability: moderate
Remaining budget: 5
Your action:
```

I Alloy Exemplar Robustness

This section reports the alloy exemplar-sensitivity analysis referenced in the main text. It shows that prompt effects in alloy-like composition editing depend not only on the existence of a demonstration, but also on the specific trajectory induced by that demonstration.

Table 34: Alloy few-shot exemplar robustness.

Condition	Success rate (%)
Zero-shot	17
Few-shot format	13
Few-shot strategy A	3
Few-shot strategy B	20

Table 35: GB1 operating-point calibration across fitness thresholds. Additive local heuristics fail across all tested thresholds, whereas the LLM and the privileged noisy policy remain above zero. This supports the claim that the chosen regime is non-trivial but non-degenerate, and that success cannot be explained by additive local improvement alone.

Policy	ft=4.7	ft=4.8	ft=5.0
Additive greedy	0	0	0
Obs. additive greedy	0	0	0
Zero-shot LLM	30	45	40
Few-shot strategy	20	20	20
Privileged noisy policy	55	55	55

J GB1 Calibration

This section provides the GB1 operating-point validation used to support the chosen threshold in the main experiments. The table shows that additive local heuristics fail across thresholds, while stronger policies remain above zero, indicating a non-trivial but non-degenerate regime.

K Word-Length Ablation

This section reports the lexical operating-point ablation over word length. Its purpose is to justify the choice of the 4-letter setting as the most informative control regime for the main study.

Table 36: Word-length ablation for lexical editing. The 4-letter setting provides the most informative operating regime: 3-letter instances are easier overall and leave less room for intervention effects, whereas 5-letter instances are sparser and harder for all methods. The larger scaffold effect at 4 letters therefore motivates its use as the main lexical setting.

Length	Zero-shot	Scaffold	Scaffold effect
3-letter	23	27	+4
4-letter	13	27	+14
5-letter	10	13	+3