

Evaluating LLM Agents Beyond Local Edit Validity

Kunhee Ryu

Yonsei University
Seoul, Republic of Korea
rgh00826@yonsei.ac.kr

Chi-Guhn Lee

University of Toronto
Toronto, Canada
chiguahn.lee@utoronto.ca

Keeheon Lee*

Yonsei University
Seoul, Republic of Korea
keeheon@yonsei.ac.kr

Abstract

Evaluations of sequential large language model (LLM) agents often report local action validity or final success. Delayed constructive tasks additionally require *recoverability*, meaning each accepted edit must leave the partial artifact completable within the remaining budget. We introduce a measurement protocol across Word Ladder, an alloy composition proxy, and the fitness landscape of the B1 domain of protein G (GB1). We call an episode surface clean when every action is well formed and locally admissible. Across six models, Word Ladder exposes interface failure with 36.1% local validity. Alloy and GB1 episodes are often surface clean, yet only 16.0% and 6.8% of them succeed. Exact enumeration of 625 GB1 variants over budgets from zero to six decomposes failure. Among surface clean episodes, 51.4% cross a recoverability boundary, while 14.1% of the remaining episodes succeed. Bounded composition audits and intervention checks connect bottlenecks to repair classes. These results establish recoverability as a decomposable and auditable evaluation axis that informs interventions for sequential agents.

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. In constructive tasks, an agent builds a solution through local edits, so a valid next action is necessary but insufficient. Even 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 therefore whether the resulting partial

solution remains recoverable, meaning that some feasible completion still exists under the remaining edit budget.

We study this gap through **constructive structure editing**, a shared evaluation substrate in which an agent incrementally modifies a partial structured output under local constraints, while final success is determined by delayed global evaluation. This framing connects structured prediction as search, iterative editing, local search, and sequential decision making under delayed feedback (Doppa et al., 2014; Guu et al., 2018; Malmi et al., 2019; Gu et al., 2019). Our contribution is an auditable evaluation methodology and reporting protocol that separates local action admissibility from trajectory state preservation. It complements existing planning theories and agent algorithms by making this separation measurable across controlled environments.

We instantiate the protocol in three controlled roles that are summarized in Figure 1. They cover lexical editing with Word Ladder, composition editing with an Alloy composition proxy, and protein sequence editing with the GB1 fitness landscape. Word Ladder serves as an action interface control because local validity is exact on the graph. The Alloy composition proxy is grounded in measured ultimate tensile strength (UTS) and density data (Borg et al., 2020), but shifts the controlled bottleneck from interface validity to delayed tradeoffs that also include recovery cost. The GB1 landscape is derived from the Fitness Landscape Inference for Proteins (FLIP) benchmark (Dallago et al., 2021) and serves as the primary exact recoverability case. Local additive evidence can be deceptive because final fitness depends on epistatic interactions. Both serve as controlled interaction proxies under the shared local edit interface. This matched design supports direct comparison while preserving each environment’s distinct delayed interaction structure and evidence scope.

*Corresponding author.

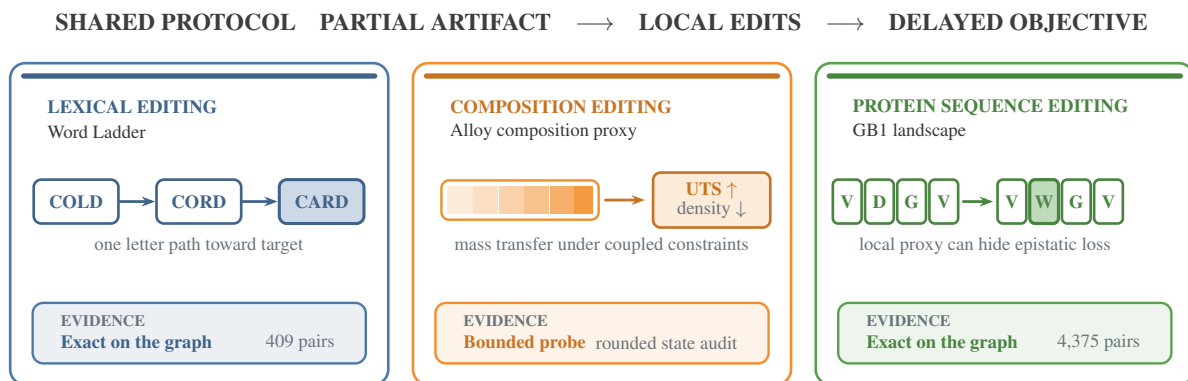


Figure 1: **Three environments, one constructive editing protocol.** Each panel follows the same partial artifact → admissible edit → delayed success flow while isolating a different bottleneck. The bottom bands distinguish exact on the graph evidence in Word Ladder and GB1 from the bounded probe and rounded state audit in Alloy composition editing. Colors follow the shared environment palette used in the remaining figures.

Formally, 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 post hoc analysis and kept hidden from the model. In GB1, exact enumeration of the restricted 625 variant landscape validates the probe and audits stored trajectories. To isolate trajectory failure from interface failure, we call an episode **surface clean** when every action is well formed and locally admissible. Failure among surface clean episodes is therefore central to our trajectory level claims. The model has satisfied the interface, but the trajectory still fails the delayed objective. Accordingly, final success alone cannot distinguish interface failure, poor stopping, local deception, and failure to find an available recovery path.

The protocol applies to tasks with four properties. These are a locally editable artifact, meaningful partial states, path dependent consequences of accepted edits, and a delayed objective that can be evaluated only after several edits. Together, these criteria specify how to extend the audit. A new environment needs a local edit graph and exact, bounded, or estimated feasible completion labels for pairs of states and budgets.

This setup yields a hierarchical result. Word Ladder first shows why surface validity must be reported, whereas Alloy composition editing and GB1 expose a stronger trajectory level failure. Many episodes are surface clean, yet success among them remains only 16.0% and 6.8%, respectively. Exact GB1 enumeration decomposes this failure rate into boundary loss and residual failure

despite retained recoverability. Once the bottleneck is localized, prompt, history, wrapper, and Monte Carlo Tree Search (MCTS) checks serve as targeted stress tests. Together, these checks and the recoverability measurements help identify 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 reporting centered on recoverability 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 editing, composition editing, and protein sequence editing. This control to mechanism ladder locates interface failure in Word Ladder, bounded recoverability evidence in Alloy composition editing, and exact recoverability auditing in GB1. Third, we show that surface clean, locally admissible trajectories often fail delayed objectives in Alloy composition editing and GB1, with GB1 failures decomposed by exact state and budget enumeration. Fourth, we use prompt, history, wrapper, and MCTS checks to show that the measured failure mechanism constrains the plausible intervention class, and we release trajectory level labels, recoverability diagnostics, exact GB1 validation, Alloy composition trajectory state audits, raw logs, and verification code.

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.

Our constructive editing abstraction is closest to structured prediction through search over structured outputs (Doppa et al., 2014), prototype based and tag based text editing (Guu et al., 2018; Malmi et al., 2019), and iterative refinement through insertion and deletion (Gu et al., 2019). It also relates to recent benchmarks that test action reasoning, asynchronous planning, and algorithm search for combinatorial optimization (Kokel et al., 2025; Gonzalez-Pumariega et al., 2025; Sun et al., 2026). Inference time methods such as consistency based sampling, 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). Our measurement setting complements this broad space of planning and search interventions by evaluating whether they preserve feasible completion in addition to emitting admissible actions.

Controlled proxies inspired by scientific design. Materials and biological sequence design involve local changes whose consequences depend on delayed properties, multiobjective tradeoffs, 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 composition editing and GB1 landscape editing as controlled sources of these delayed interactions. The environments provide diagnostic tests of feasible completion preservation under a simple local edit grammar, with scientific grounding that motivates the delayed interactions.

Across these lines of work, our measurement oriented reframing makes recoverability observable and, in GB1, decomposable. This trajectory level quantity extends local action validity for sequential agent evaluation.

3 Method

Constructive editing centered on recoverability

For each environment e , define a task $\mathcal{T}_e = (\mathcal{S}_e, \mathcal{A}_e, P_e, O_e, G_e, H_e)$, where $\mathcal{S}_e$ is the state space, $\mathcal{A}_e = \mathcal{A}_e^{\text{edit}} \cup \{\text{FINALIZE}\}$ is the action set, P_e , O_e , and G_e are the edit transition, observation, and delayed evaluation functions, respectively, and H_e is the edit budget. 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 $G_e(\tau_{\leq t}, x_t)$. Writing τ for the full trajectory, terminal success is $S(\tau, x_T) \in \{0, 1\}$ and may depend on both the final state and trajectory induced quantities. 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 the rendered task state and local feedback, while the full environment state and post hoc recoverability labels remain hidden. The term *constructive editing* names the edit substrate on which we define the recoverability measurements below. In particular, the action interface and trajectory state parts of the decision problem are logged separately, so they can fail separately.

We distinguish five 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. We use it as a task specific post hoc diagnostic that remains hidden from 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 the event that an episode is surface clean, let S abbreviate the terminal success event $S(\tau, x_T) = 1$, and let B denote an exact recoverability boundary crossing before termination. Because B precludes success,

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

Section 5.3 applies this decomposition to the six model GB1 audit. Among surface clean episodes, 51.4% cross this exact boundary and 14.1% of the remaining episodes succeed. Together these factors yield the reported 6.8% success among surface clean episodes. For Alloy composition editing, recoverability drops provide bounded probe degradation signals at the analogous boundary.

Environments

Word Ladder 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 loss of reachability within the remaining budget are directly visible before the trajectory level claims in the other environments.

Alloy 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 a grounded estimate of ultimate tensile strength (UTS) (Borg et al., 2020), exact density, and a path dependent recovery cost that increases when the composition moves away from a known valid region. The environment is a controlled compositional optimization proxy with grounded data and surrogate components. Its diagnostic focus is trajectory management. A locally attractive edit can improve one property while worsening the bounded recoverability estimate or making successful completion require nonmyopic recovery edits.

GB1 landscape editing. The state is a restricted sequence over selected GB1 positions. An edit changes the amino acid at one editable position. Intermediate feedback includes local additive estimates, but final success depends on an epistatic landscape grounded in the GB1 measurements used by the Fitness Landscape Inference for Proteins (FLIP) benchmark (Wu et al., 2016; Dallago et al., 2021). The environment is a restricted deceptive landscape stress test for exact recoverability auditing. 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. The conditions are zero shot, format only few shot, strategy based few shot, scaffold, generic self check, and task specific self check. We interpret prompt effects as intervention class diagnostics under fixed templates. 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 arise beyond the action interface. A sequence of locally admissible edits eliminates a feasible completion or degrades the recoverability estimate, violates delayed constraints, ends in premature finalization, or falls into a local optimum. Recoverability drops and local deception events provide mechanism evidence beyond terminal failure labels. In GB1 these diagnostics are checked against exact enumeration. In Alloy composition editing they remain bounded probe evidence.

4 Experimental Setup

The main controlled sweep uses Qwen2.5-7B-Instruct-Turbo with deterministic hosted model inference, fixed prompt templates, and 50 episodes for each environment and prompt cell, using seeds from 0 to 49. 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 binomial success rates in the headline table. Appendix B.9 adds an episode level bootstrap audit, a matched cluster bootstrap defined by model and prompt, and a fixed effect logistic robustness check over a separate six model raw log audit. These uncertainty estimates quantify descriptive variation across the evaluated episodes.

The release contains raw per episode logs, machine readable summaries, verification code, baseline calibrations, operating point checks, ablations, cross model analyses, and search pilots. Code and curated results are publicly available at <https://github.com/rgh112/llm-agents-beyond-local-edit-validity>. Cross model results test qualitative recurrence and model conditional effects. Appendix A specifies

the grounded and controlled components and the resulting claim scope for each environment.

Baseline hierarchies establish operating points between pure local optimization and degenerate impossibility. For GB1, we additionally compute an exact state and budget recoverability table over the restricted landscape. For Alloy composition 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 nontrivial feasible completions exist. This hierarchy establishes task solvability and supports a trajectory construction interpretation of low LLM success.

5 Results

We begin with cross model aggregates because they establish the phenomenon of surface clean failure 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 mechanism ladder

We first separate interface failure from trajectory level failure. Table 1 aggregates the six model sweep over three prompts and 50 seeds per cell. It reports unconditional success as well as success among surface clean episodes, making the size and selectivity of the clean subset explicit. Word Ladder is the control case. Local edit validity is only 36.1%, so final success partly reflects action surface failure. Alloy composition and GB1 editing shift the bottleneck. In Alloy composition editing, the surface clean rate is 83.8%, but success among surface clean episodes is only 16.0%. In GB1, the surface clean rate is 84.4%, local edit validity is 92.7%, and surface clean success is only 6.8%. Across Alloy composition and GB1 editing, 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 matched protocol reveals a ladder of distinct bottlenecks. Word Ladder diagnoses the action interface, Alloy composition editing diagnoses delayed tradeoffs 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. This pattern consists of valid moves along paths that cannot reach the target within the remaining budget. In Alloy composition and GB1 editing, trajectory level structural events dominate the event mix. They account for 97.5% of all classified surface and structural failure events in Alloy composition editing and 85.3% in GB1. Alloy composition editing is dominated by OBJECTIVE TRADEOFF failure, where local improvements worsen the tradeoff between strength and density 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 form a deterministic diagnostic set specific to the suite. Their value is that they align with the intended bottleneck of each environment. They cover reachability within the remaining budget, delayed multiobjective 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 computed post hoc and remain hidden from the model.

The 0.0% Word Ladder finalize rate reflects automatic target reach termination. 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.

Exact GB1 enumeration covers all 625 variants at budgets from 0 to 6. This gives 4,375 pairs of states and budgets with zero probe mismatches. Across all six models, the 6.8% surface clean success rate factors into a 51.4% boundary crossing

Table 1: Cross model outcomes and recoverability evidence hierarchy (900 episodes per environment). Surface cleanliness is the episode level rate with no malformed or locally invalid action. Success is unconditional. Success | clean is conditioned on surface cleanliness. “Exact” refers to the stated finite graph. Continuous real world design spaces remain outside this scope.

Environment	Local valid	Surface clean	Success	Success clean	Recoverability evidence
Word Ladder	36.1%	24.1%	33.8%	75.1%	Exact reachability on the graph with 409 pair instances, enumerable by BFS
Alloy composition	67.6%	83.8%	16.2%	16.0%	Bounded online probe, exact only on the implemented rounded state graph
GB1	92.7%	84.4%	6.6%	6.8%	Exact on the graph for 4,375 state and budget pairs with zero probe mismatches

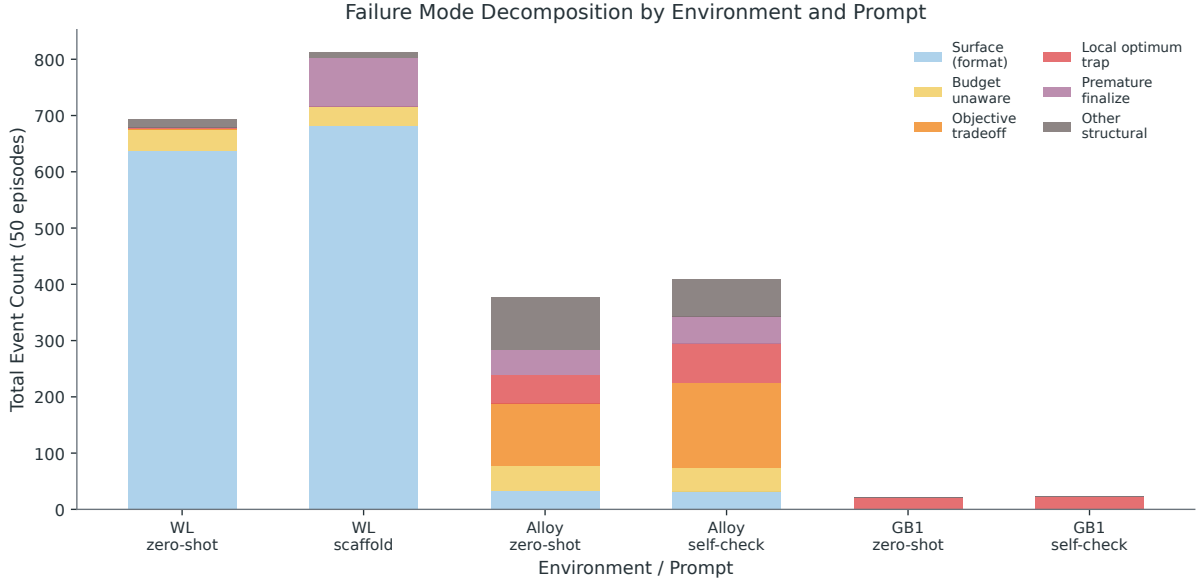


Figure 2: Failure composition across environments. Word Ladder acts as an interface and budget control. Alloy composition and GB1 editing show many trajectory level structural failures even when surface failures are limited, locating their dominant bottleneck in delayed construction.

Table 2: Trajectory diagnostics over an additional nine model diagnostic scope subset, separate from the six model exact GB1 raw log audit summarized 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 composition editing, a recoverability drop denotes a bounded probe score decrease. Beam certified binary infeasibility is evaluated separately.

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

rate and 14.1% residual boundary preserving success. Table 3 shows what these labels add beyond validity and final success. Gemma, Mistral, and Qwen have identical 100% local validity and surface cleanliness and all rarely succeed, yet exact

Table 3: Selected GB1 model level results over 150 episodes per model. Boundary crossing is exact among surface clean episodes. Brackets are Wilson 95% CIs. The three models with 100% validity are nearly indistinguishable by validity and success but sharply separated by recoverability.

Model	Local valid	Clean	Success	Exact crossing
Gemma-3-27B	100.0	100.0	6.0	5.3 [2.7, 10.2]
Mistral-24B	100.0	100.0	2.0	41.3 [33.8, 49.3]
Qwen3-32B	100.0	100.0	0.7	62.7 [54.7, 70.0]
Llama-3.3-70B	99.9	89.3	4.7	82.1 [74.7, 87.7]

crossing ranges from 5.3% to 62.7%. The cheap estimated drop rates for these models are 0.05, 0.41, and 0.63, closely matching the exact ordering. Llama-3.3-70B has the highest crossing rate in this audit (82.1%).

Worked trajectory. In Qwen3-32B seed 0, six admissible edits repeatedly raise the visible addi-

tive estimate to 3.82 while oscillating between edits at D40 and V39. The trajectory never mutates G41 (editable index 2), although 40 of the 44 successful variants require a mutation there. True fitness remains 2.59, below the 5.0 success threshold. At edit 6 the visible proxy rises by 0.38 while the exact label for the current state and budget flips from recoverable to unrecoverable. Several mechanisms could produce this observation, including a misaligned visible objective, insufficient exploration, or the stopping policy.

Across the landscape, additive and true fitness correlate only $r = 0.20$, with zero top five and top ten overlap. Visible proxy MCTS reaches 74% success whereas privileged MCTS reaches 100%. Conversely, 133/141 surface clean Gemma failures never cross the boundary. They retain recoverability but are not found by search.

Alloy composition evidence is deliberately weaker than GB1’s exact full graph audit but complementary, and its two audits answer different questions. The bounded beam audit asks whether failed surface clean trajectories can be certified impossible by a visible state completion search. Its absence of beam certified losses supports interpreting bounded probe drops as degradation signals. The implemented state exact audits track the rounded simulator state, including path dependent recovery cost, and show that many logged trajectories cross an implemented state boundary.

A broad implemented state audit that allows up to 10,000 expansions per state 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 higher cap slice over visited states with at most four edits remaining 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 composition therefore provides complementary evidence at two scopes. Some failures look recoverable under bounded visible search, while many logged trajectories cross exact boundaries in the implemented rounded state graph.

5.4 Recoverability measurements make intervention targets testable

The practical use of the measurement protocol is to connect trajectory labels to candidate intervention hypotheses. Table 4 summarizes this mapping. The same final success rate can hide action interface er-

rors, poor stopping, local deception, or failure to find an available recovery path. The trajectory diagnostics separate these cases. Our prediction language is operational. The measurements identify the next repair class to test.

Existing cost adjusted wrapper checks show partial, environment dependent gains. In the open family shallow wrapper evaluation (Appendix C.13, Table 15), Alloy composition editing improves from 6.7% greedy success (8.0 calls/episode) to 22.0% under consistency based sampling (39.6 calls/episode), while Word Ladder improves more modestly (48.7% to 54.7%) and GB1 stays flat or worse (20.7% to 18.0%). Appendix C.12 reports a depth 3 visible state beam stress test. It leaves GB1 at success rates from 8% to 18% across three open family small models, with substantial failure remaining under sampled lookahead.

The scorer ablation gives the same message. Switching beam scoring from a visible state scorer to a privileged scorer raises Alloy composition success from 11.7% to 21.7%, while Word Ladder and GB1 show nonmonotonic changes. The benefit of privileged target information therefore concentrates where the diagnostic identifies delayed tradeoff management as the bottleneck. The same appendix adds a non LLM MCTS based boundary check. Visible proxy MCTS solves 74.0% of GB1 but only 6.0% of Alloy composition seeds, while privileged terminal objective MCTS solves all GB1 and 82.0% of Alloy composition seeds. The aggregate wrapper, depth 3 beam, and scorer ablation summaries preserve the same ranking. Alloy composition gains most from visible search and sampling, GB1 visible wrapper deltas remain negative, and Word Ladder gains are mainly interface side under higher call budgets.

The pattern is consistent with this hypothesis. Alloy composition benefits from search based sampling that helps manage delayed tradeoffs, whereas GB1 remains limited unless the intervention exposes the objective relevant to recoverability or adds a stronger backtracking policy. We report calls per episode here and token usage in the released summaries where hosted APIs provide reliable counts.

5.5 Format examples yield limited and inconsistent gains

Table 5 reports the main controlled prompt sweep. If weak performance were primarily a schema compliance problem, format only few shot demonstra-

Table 4: Recoverability measurements as intervention hypotheses. The rows form a suite specific map from trajectory labels to the next repair class to test.

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

tions should improve broadly. In this sweep, format only prompting falls below zero shot in Word Ladder, Alloy composition editing, and GB1. The best intervention also differs by environment. Word Ladder benefits most from a scaffold, while Alloy composition and GB1 editing benefit most from task specific self checking.

We interpret prompts as lightweight policy guidance interventions. The observed pattern is consistent with the diagnosed bottlenecks. A scaffold explicitly supports reachability and budget tracking in Word Ladder, while task specific self checking directs attention to misleading local evidence in GB1. Alloy composition editing remains low across prompting conditions, showing that the tradeoff between strength and density requires more than additional format exposure.

5.6 Raw history has model and task conditional effects

Appendix F shows model and task conditional effects of recent raw edit history. In the controlled Qwen history ablation, Alloy composition zero shot improves from 12% with 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 history ablation row drops from 35% to 5%. The narrower conclusion is therefore diagnostic. Raw trajectory context can either support recoverability tracking or reinforce earlier poor commitments.

5.7 Cross model checks show heterogeneous recurrence

The six model sanity check and 13 model open family panel show qualitative recurrence and model dependent performance. Across the six model grid, format only exemplars remain inconsistent and

structural events dominate surface events in Alloy composition and GB1 editing. The larger 13 model panel under the structural prompt 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 composition editing, and Ministral-8B leads GB1. The top model therefore varies across environments.

This model dependence is part of the protocol output. A useful measurement protocol should separate models and interventions in ways that reflect different constructive bottlenecks. Stability and sensitivity checks in Appendix C.13 show that scorer, decoding, and hosted model API perturbations move some cell level rates, while memory, prompt, and surrogate choices produce model specific responses. We therefore state recurrence claims qualitatively. Tested models often show recoverability degradation despite locally admissible edits, with exact validation available for GB1. Exact success rates and best interventions remain model conditional.

An NLP native document plan bridge tests the same edit abstraction on a five section outline with local structural constraints and a delayed global plan constraint (Appendix C.11). A scaffold raises GPT-4o-mini success from 44% to 72%, while GPT-4o is near saturation (98% to 94%) and open family effects are mixed. This diagnostic extension shows that the protocol can expose model conditional editing bottlenecks in an NLP native artifact. Broader naturalistic generalization remains future work.

5.8 Calibration establishes informative operating points

Baseline and calibration checks support the operating points. Word Ladder requires target reacha-

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 yield limited and inconsistent gains, 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 (four letter)	18 [10,31]	8 [3,19]	18 [10,31]	30 [19,44]	6 [2,16]	18 [10,31]
Alloy composition	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]

bility to be preserved within the remaining budget in addition to repeated valid edits. Alloy composition editing couples local property improvement with density and recovery cost. Greedy edits can improve estimated strength while worsening the other objectives. GB1 provides the sharpest check against a purely additive explanation. Additive greedy baselines fail across tested thresholds, while LLM and privileged noisy policies remain above zero. This separates successful trajectories from both chance level behavior and additive local improvement.

Appendix B.9 adds further robustness checks. On the six model raw logs, bootstrap intervals preserve the same ordering. Success among surface clean episodes is high in Word Ladder at 75.1% with a 95% CI from 69.6 to 80.2. It is low in Alloy composition editing at 16.0% with a CI from 13.3 to 18.8 and in GB1 at 6.8% with a CI from 5.3 to 8.7. Paired cluster bootstrap and fixed effect logistic checks preserve this separation. An all episode logistic model that includes surface clean subset membership and its environment interaction yields the same adjusted clean rates (73.9%, 15.5%, and 6.3%), preserving the environment ranking after conditioning. Small non LLM beam pilots also solve audited starts in Alloy composition editing and GB1 under the same edit grammar. Together, these checks attribute low LLM success to trajectory construction difficulty while establishing attainable solutions.

6 Discussion

The main implication is that recoverability provides a measurable trajectory level quantity throughout an episode. Word Ladder shows that local validity itself can be limiting. Alloy composition editing moves beyond the interface to delayed trade-off management. GB1 shows exact recoverability loss under deceptive local evidence. The resulting claims concern interaction structure within these

controlled proxies.

Table 4 is the operational form of the claim. Gains from format examples, history, shallow wrappers, and MCTS based search concentrate when the intervention targets the measured bottleneck. Sequential LLM evaluations should therefore report validity, surface clean success, recoverability loss, local deception, finalize behavior, dominant failure class, and intervention cost as separate axes alongside aggregate success.

7 Conclusion

We introduced a measurement protocol centered on recoverability for sequential construction over partial structured outputs. The central finding is a separation of bottlenecks under a matched edit substrate. Word Ladder exposes action interface failure. Alloy composition editing exposes a delayed tradeoff and trajectory management burden. GB1 shows that surface clean failures can be decomposed into exact recoverability boundary loss and residual failure despite retained recoverability. Recoverability is therefore a reporting axis for sequential LLM agents that is measurable, auditable, and informative for interventions. It shows when valid local actions cease to preserve feasible completion and which class of repair is plausible.

Limitations

Scope of the constructive abstraction. Our environments abstract much of real scientific design. Alloy development depends on thermodynamics, kinetics, processing, and microstructure, while biological sequence design involves structural, functional, and experimental constraints beyond the GB1 proxy. We therefore evaluate *constructive editing structure* rather than full domain realism.

Gap between surrogates and reality. Alloy composition and GB1 use surrogate or semisynthetic scorers for controlled large scale evaluation.

Deployment claims require full simulation or laboratory validation.

Restricted edit grammars. Constrained edit grammars isolate trajectory management and improve interpretability, but underrepresent open design spaces.

Scale of edit horizons. Modest edit budgets support reproducible hosted model evaluation and induce path dependence, but remain shorter than many design workflows.

Recoverability validation. For GB1, exact enumeration validates the probe over 625 variants and budgets from 0 to 6. Alloy composition has two narrower audits (Appendix B.9). A visible state bounded beam certifies none of the audited surface clean failures as infeasible, while finite state audits with latent recovery cost identify exact loss in the implemented graph. Thus bounded probe drops indicate tradeoff and lookahead burden, while rounded state losses are exact only within the simulator. Continuous feasibility remains untested, and an exact all budget table would need composition, phase, budget, and recovery cost.

Planner and statistical baselines. Non LLM beam pilots establish feasible audited starts, and MCTS separates visible proxy search from privileged terminal search. These are boundary checks rather than a full planning study with learned values, rollback policies, and richer information assumptions. Bootstrap intervals quantify episode uncertainty, while the paired cluster bootstrap and fixed effect logistic adjustment cover selected robustness axes. A full mixed effects analysis remains future work.

Search wrapper scope. Our main experiments use action only editing with prompt and history interventions. Additional stress tests include cost adjusted and consistency based sampling, visible state beams, loop avoidance, and an MCTS boundary check (Appendix C.12). Privileged MCTS provides a diagnostic upper bound through explicit action enumeration and terminal information. Learned values, rollback, and tool assisted refinement remain future work.

Exemplar sensitivity and model scope. Strategy based Alloy demonstrations are sensitive to the selected trajectory. The main controlled study uses one hosted model, strengthening within model

comparison but limiting provider generalization. A cross model sanity check, 13 model panel, and shallow wrapper evaluation broaden coverage over an intentionally unbalanced grid (§5.7, Appendix C.13). These results support controlled trajectory management claims, while scaling claims require balanced coverage across families and sizes.

Ethical Considerations

Public lexical, alloy property, and protein fitness landscape data are used. No human subjects, private user data, annotation labor, or deployment decisions are involved. The diagnostic proxies contain no laboratory, synthesis, DNA design, or biosynthetic guidance. The primary risk is mistaking proxy success for deployable scientific capability. We mitigate it by separating grounded and controlled components, reporting claim boundaries, and releasing raw logs and verification code. Hosted model APIs are used only for evaluation. The repository contains no secrets and requires no hosted calls.

Acknowledgments

This research was supported by the Basic Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (2710108964), in part by a grant from the Institute for AI and Social Innovation at Yonsei University, in part by the Yonsei University Office of Research Affairs through the Yonsei T.R.U.S.T. (Yonsei Transdisciplinary Research Union for a Sustainable Tomorrow) program (Project Y, project no. 2025-22-0461), and in part by the Institute of Convergence Science (ICONS), Yonsei University.

AI assistants were used for language polishing and coding support. All scientific claims, experimental designs, analyses, and final manuscript content were reviewed and verified by the authors.

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, Lilian Weng, and Aleksander Madry. 2025. [MLE-bench: Evaluating machine learning agents on machine learning engineering](#). In *International Conference on Learning Representations*.
- Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. 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*, volume 36, pages 28091–28114.
- Janardhan Rao Doppa, Alan Fern, and Prasad Tadepalli. 2014. [Structured prediction via output space search](#). *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. [M²hub: Unlocking the potential of machine learning for materials discovery](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Gonzalo Gonzalez-Pumariiega, 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*, volume 32, 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 Wiegrefe, 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*, volume 36.
- 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. Kaczmarek, 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*, volume 36.
- 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*.
- Nicholas C. Wu, Lei Dai, C. Anders Olson, James O. Lloyd-Smith, and Ren Sun. 2016. [Adaptation in protein fitness landscapes is facilitated by indirect paths](#). *eLife*, 5:e16965.
- 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*, volume 37, pages 52040–52094.
- Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. [Webshop: Towards scalable real-world web interaction with grounded language agents](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757.
- Shunyu Yao, 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*, volume 36, pages 11809–11822.
- 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 scope for the three environments. The suite targets constructive editing structure across lexical, composition, and protein-sequence settings.

Environment	Grounded component	Controlled component	Claim scope
Word Ladder	Enable1 lexicon and Hamming-1 word graph	sampled start–target pairs, edit budget, hidden shortest-path distance	lexical path construction
Alloy composition editing	Multi-Principal Element Alloy (MPEA) measured-property data (Borg et al., 2020) and exact density calculation	heavy-element scaling, recovery cost, thresholds, observation noise	controlled composition trajectory management
GB1 landscape editing	Fitness Landscape Inference for Proteins (FLIP) GB1 fitness landscape over measured variants (Dallago et al., 2021)	four editable sites, noisy additive observations, stability gate	restricted sequence landscape trajectory management

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, with *constructive structure editing* as the operational label for the matched task family.

B.1 Constructive Editing Task

For environment e , a constructive editing task is

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

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

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

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

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

while FINALIZE triggers delayed evaluation

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

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 measured or estimated recoverability. 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 (Doppa et al., 2014). 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 a terminal event: 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

Word Ladder 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 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. Its claim scope is controlled composition-editing trajectory management.

GB1 landscape editing. The partial solution is a restricted sequence variant. A locally admissible edit changes the amino acid at one selected position. Feasible completion corresponds to remaining able to reach a successful variant on the restricted epistatic landscape. Its claim scope is recoverability within the restricted protein-sequence landscape.

B.4 Prompted Policy and History Conditions

A model m 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

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

Prompt conditions are fixed intervention classes shared across models. History conditions vary access to raw recent trajectory information: $M0$ exposes only the current rendered state, while $M1$ and $M3$ expose one or three recent edits. These conditions isolate raw history access within the benchmark.

B.5 Prompt and History Effects

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

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

For history condition k , we define

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

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 occur despite valid interface use and prevent appropriate termination or state progression, such as premature finalization or repeated attempts that do not advance the state.

Trajectory level structural failures. These are failures arising beyond the action interface. Examples include budget-unaware continuation in Word Ladder, objective-tradeoff collapse in Alloy composition 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 task specific diagnostics are computed over stored trajectories and kept hidden from the model. We use an exact-feasibility interpretation only where it is separately validated for the environment.

B.8 Recoverability Probe Scope

The reported recoverability probes are environment specific diagnostics. In Word Ladder, the fixed graph permits exact reachability checks under the remaining budget. 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 boundary preserving episodes succeed; equivalently, 391 of 708 surface clean failures cross the exact boundary before finalization. In Alloy composition editing, the discrete edit grid makes exhaustive or large-beam feasibility analysis plausible under fixed delta sets, but the online label remains a bounded diagnostic probe. Section B.9 separately audits visible state beam reachability and exact reachability on the implemented rounded simulator graph. Thus an online Alloy composition recoverability drop denotes bounded probe degradation, while the separate rounded state audit determines exact transitions within the implemented graph.

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 composition Beam Audit, and Non-LLM Search

We add six lightweight audits of diagnostic soundness. First, an episode level bootstrap over the six model raw logs confirms recurrence of the headline pattern across cells: surface clean success remains high for Word Ladder but low for Alloy composition and GB1. Second, a paired cluster bootstrap over matched model–prompt cells preserves the ordering across model families and prompt classes. Third, a grouped-binomial fixed effect logistic check adjusts for model and prompt-class indicators and preserves the environment ordering. Fourth, a second grouped model retains all episodes and includes surface clean subset membership plus environment interactions as covariates; this checks whether conditioning itself creates the separation. Fifth, for Alloy composition editing, we run three trajectory state audits: a depth-capped visible state beam over the fixed edit grid, a search capped at 10,000 expansions over all visited rounded states, and a higher cap search over visited states with at most four edits remaining. The visible state beam tests whether failed trajectories remain recoverable under explicit search, while the implemented state searches can include the latent recovery cost state; none is a continuous-state feasibility proof. Sixth, 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 the existence of feasible completions under simple search; the LLM evaluations separately measure one-action-per-call discovery.

B.10 Baseline Hierarchy

The baseline hierarchy checks operating point validity across distinct policy classes. 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 nontrivial feasible completions exist. A useful environment places greedy local optimization below policies with access to the relevant trajectory level structure.

C Environment Construction Details

This appendix describes how the three evaluation environments were constructed and calibrated. We instantiate a shared constructive editing protocol in which local edits, path dependence, and delayed global evaluation are all consequential, prioritizing controlled comparison over domain realism. 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 many nonterminal states the agent faces multiple valid actions, many of which lead to suboptimal or failed final states. Surface formatting correctness therefore does not by itself predict task success.

Table 8: Soundness audits supplementing final success rates. Bootstrap CIs use six model raw logs. The Alloy composition beam audit is restricted to unique visited trajectory states in the bounded visible state 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 composition failure shows beam certified loss. The all-visited implemented state audit uses a cap of 10,000 expansions per state. The higher cap exact slice uses a cap of 250,000 expansions over visited states with remaining budget at most four. In both audits, capped states are reported as unknown. Slash-separated rates follow Word Ladder / Alloy composition / GB1 order. Non-LLM beam pilots serve as operating point checks.

Audit	Unit	Result
Word Ladder surface clean success	bootstrap 95% CI	75.1 [69.6, 80.2]
Alloy composition 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 composition surface clean episodes	episodes	754/900
GB1 surface clean episodes	episodes	760/900
Word Ladder–Alloy composition 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 composition–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 composition 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 composition vs. GB1 adjusted odds ratio	fixed effect logit	2.641 [1.052, 6.629]
Adjusted clean success, retaining all episodes	selection-adjusted logit	73.9 / 15.5 / 6.3%
Alloy composition vs. Word Ladder clean odds ratio	selection-adjusted logit	0.060 [0.026, 0.141]
GB1 vs. Word Ladder clean odds ratio	selection-adjusted logit	0.022 [0.009, 0.053]
Alloy composition beam/probe agreement	visited states	54/54
Alloy composition surface clean failures with beam certified loss	episodes	0/633
Alloy composition all-visited states exact-labeled	visited states	5477/6143
Alloy composition all-visited exact/probe agreement	known visited states	5449/5477
Alloy composition all-visited failures with exact loss	surface clean failures	614/633
Alloy composition higher cap implemented states exact-labeled	budget ≤ 4 states	4166/4193
Alloy composition higher cap exact/probe agreement	known budget ≤ 4 states	4084/4166
Alloy composition failures with implemented exact loss	surface clean failures	607/633
GB1 failures crossing exact boundary	surface clean failures	391/708
GB1 residual exact-recoverable failures	surface clean failures	317/708
Alloy composition visible beam pilot	starts	10/10
Alloy composition privileged beam pilot	starts	10/10
GB1 visible beam pilot	starts	20/20
GB1 privileged beam pilot	starts	20/20

3. **Baseline separability.** The operating regime was chosen so that random, observation-level greedy, and privileged diagnostic baselines occupy distinct performance tiers.
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 while logging any failure events for post hoc analysis.
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.** At termination, the environment evaluates success against its hidden global objective.

The edit grammar is environment specific, but every edit action 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 composition	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 \Delta $	4×4
Budget	$5 \times k_{\text{opt}}$	8 steps	6 steps
Instance source	409 pairs	seed-jittered starts	625 variant landscape
Observation noise	none	strength ± 30 MPa	fitness ± 0.20
Key difficulty	interface validity and reachability	strength–density tradeoff	epistasis trap

C.3 Word Ladder Instance Construction

Lexicon and graph. We use the Enable1 English word list¹. After filtering to four letter words, we obtain a vocabulary of 3,903 words. We construct a Hamming-1 adjacency graph over this vocabulary ($|E| = 21,002$); the largest connected component contains 3,807 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 four 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 composition 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.

¹<https://norvig.com/ngrams/enable1.txt>

Ultimate tensile strength surrogate. We train a gradient-boosting regressor on 539 usable records with positive reported UTS 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.

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

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

where $h = 1.0 \cdot c_{\text{Mo}} + 0.6 \cdot c_{\text{Co}} + 0.3 \cdot c_{\text{Ni}}$, $\alpha = 0.30$, and $h_{\text{thresh}} = 18.0$. 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 composition proxy 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 when heavy-element scaling and recovery cost constraints are removed (100%), and drops when both are tightened (55%). This shows that the environment requires coupled tradeoff management while retaining high privileged-policy solvability. The LLM sensitivity evidence comes from 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 track Euclidean distance from the reference composition (Fe:20, Ni:20, Cr:25, Co:10, Mn:10, Mo:15). Moving farther away by Δd adds $0.22\Delta d$ to recovery cost; moving closer removes 30% of that corresponding amount, with a floor at zero. If recovery cost reaches or 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 (Gaussian noise with standard deviation 30 MPa), 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: $\text{UTS}_{\text{eff}} \geq 2,000$ MPa, density ≤ 8.20 g/cm³, 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 (Wu et al., 2016; Dallago et al., 2021). Each position admits five residues, yielding a combinatorial space of $5^4 = 625$ variants. The wild-type sequence is VDG V. We index these editable sites from 0 to 3, so position 2 refers to G41.

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 show structural misalignment between local additive evidence and the delayed epistatic objective beyond observation noise.

Stability mechanic. To create path dependence, we assign each mutation a controlled 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 present in 40 of the 44 variants above the success threshold.

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 (Gaussian noise with standard deviation 0.20) 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 yield low random and myopic performance, higher privileged-policy performance, and room for prompt-induced variation among LLM agents.

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 composition. 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 a fitness threshold of 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 composition. The starting-composition distribution is continuous, so there is no fixed instance pool. 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.

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. Baseline rows use seeds 0–99; LLM rows use 50 episodes per prompt cell. The intended calibration regime is visible in the separation between these tiers. Word Ladder and Alloy composition editing admit useful but limited myopic heuristics; GB1 instead makes additive myopic reasoning fail, while privileged epistasis-aware policies remain high.

	Word Ladder	Alloy composition	GB1 landscape
Instance pool size	409 pairs	seed-jittered starts	625 variants
Budget (steps)	$5k_{\text{opt}}$ (15–30)	8	6
Avg. branching factor	9.9	$\sim 30/\text{step}$	$\sim 4/\text{step}$
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%

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.

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 controlled LLM sweeps used Qwen2.5-7B-Instruct-Turbo through the Together hosted model API. Cross model sanity checks queried OpenRouter-hosted instruction models with temperature 0.0 and identical environment implementations. The released code includes raw per episode logs, machine readable aggregate summaries, environment implementations, prompt templates, task instances, and source data needed to reproduce all reported aggregate tables without unpublished intermediate files. No model training is performed. Hosted experiments are inference-only and are reported with episode counts, model-call budgets, and released token summaries; the non-hosted verification and table-generation code operates on stored logs and requires neither GPUs nor external API calls. The released code is licensed under the MIT License; source datasets retain their original terms. The GB1 environment is a restricted sequence landscape proxy and contains no laboratory protocol, synthesis instruction, DNA design workflow, or biosynthetic design guidance. The released verification code checks consistency between raw logs and reported aggregates and also verifies the deterministic Alloy composition, GB1, and document plan bridge diagnostics. The document plan diagnostic makes an NLP native bridge task auditable. OpenAI and open family bridge runs provide additional evidence around the main three-environment recoverability measurement protocol.

C.10 Trajectory Diagnostic Definitions

For every non-final accepted edit $a_t : x_t \rightarrow x_{t+1}$, let b_t be the pre-edit budget, so the post-edit budget is $b_t - 1$. We record an environment specific future-feasibility probe $R_e(x, b)$. Higher R_e is defined to indicate stronger evidence that the state can be repaired within the remaining budget. Word Ladder uses normalized shortest-path slack to the target; Alloy composition 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

$$R_e(x_{t+1}, b_t - 1) < R_e(x_t, b_t) \quad (10)$$

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 examine the constructive editing framing in an additional NLP native setting, we include a small document plan editing task in which a model edits a five section outline under local structural constraints and a global target-plan constraint. This task serves as a bridge probe outside the main three-environment aggregate.

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, and the scaffold’s main effect is a reduction in surface events. Together, these results already show model dependence; the open family table below tests whether the same task is informative for smaller open family models.

Table 13: Open-family document plan bridge results. 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 shows model specific effects: Qwen3 succeeds most often under zero shot prompting, Llama benefits from the scaffold, and Ministral remains at 0/50 under both prompts despite high locally valid edit rates. This mixed result provides boundary evidence for the bridge while preserving the paper’s main claim on the matched Word Ladder, Alloy composition, and GB1 constructive editing protocol.

C.12 Single-Model Stress-Test Pilots

We also ran small OpenAI pilots for plausible alternatives and sensitivities: independent restarts, shallow search wrappers, surrogate sensitivity, memory controls, and date-separated hosted model API stability. Their small-sample, mostly single-model scope makes them sensitivity analyses that complement the main evidence and make the findings more auditable.

The GPT-4o-mini wrapper pilots show heterogeneous effects from extra calls: local validity improves in some cells and Alloy composition occasionally benefits, while Word Ladder consistency based sampling 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 composition/GB1, and GB1 remains unsolved in this small sample. A targeted GPT-4o wrapper pilot over the same three environments, seeds 0–4, visible state scoring, $k = 3$, and beam depth 1 gives a stronger-model contrast: Word Ladder reaches 2/5 greedy, 3/5 consistency based sampling, 1/5 beam, and 3/5 loop-avoidant; Alloy composition 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, while the GB1 landscape failure persists in this pilot.

We also ran an open family shallow wrapper evaluation 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 results and provenance

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

Pilot	Environment	Success	Calls/episode	Structural
Restarts	Alloy composition	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 composition	0/5	8.0	44
Self-consistency	Alloy composition	1/5	24.0	33
Beam	Alloy composition	1/5	24.0	40
Loop-avoidant	Alloy composition	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

details are provided with the released code. The clean evaluation subset excludes three quarantined network/API-error logs, so Qwen Word Ladder beam has 49 clean seeds and Qwen Alloy composition greedy has 48 clean seeds. Overall, shallow wrapper gains are environment- and model dependent: they help Llama Alloy composition substantially and sometimes improve Word Ladder, while GB1 remains governed by local-optimum traps even when local validity is near perfect.

Table 15: Open-family shallow wrapper evaluation. Values are success rates over 50 seeds unless marked clean- N after quarantining network/API-error raw logs. Best wrapper excludes deterministic greedy; “self” denotes consistency based sampling and “loop” denotes loop avoidance. 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 composition	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 composition	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 composition	6%	34% self	38.8
Llama-8B	GB1	22%	26% loop	27.1

The same evaluation reinforces the cost-sensitivity of any wrapper claim. For example, Llama Alloy composition consistency based sampling 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 shallow visible state wrappers partially close the structural gap, with uneven gains and substantial added cost across environments.

We then ran a costlier depth 3 visible state beam stress test for the same three open family small models. The planner samples candidate actions from the model, scores visible next states, and expands a depth-limited tree using no exact recoverability information, rollback, or learned value function. Table 16 shows heterogeneous effects. It raises some Word Ladder and Alloy composition cells, but leaves GB1 at 8–18% success with LOCAL OPTIMUM TRAP still dominant. The result points to trajectory level information or backtracking as the missing capability in GB1, beyond additional sampled visible state lookahead.

To separate more sampled LLM continuations from explicit search architecture, we also ran a non LLM MCTS based 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

Table 16: Open-family depth 3 visible state beam stress test. All rows use the structural prompt, state only memory, visible state 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 composition	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 composition	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 composition	10/50	9/48	151.0
Llama-8B	GB1	9/50	9/49	85.0

solves many GB1 seeds but still leaves 13/50 local-optimum failures, locating the residual bottleneck beyond action formatting. Second, privileged MCTS solves GB1 completely and substantially improves Alloy composition, showing that explicit state search with the right objective information can change the intervention profile. Together, these results establish planner solvability while localizing the information gap in the LLM wrappers.

Table 17: Non-LLM MCTS based boundary check over seeds 0–49. Visible mode uses visible proxy rollout value; privileged mode uses terminal objective information as a diagnostic upper bound. Each decision uses 64 MCTS simulations.

Environment	Mode	Success	Mean steps	Dominant residual failure
Alloy composition	Visible proxy	3/50	7.58	hard constraint
Alloy composition	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 pilot runs with shorter output limits produced empty actions, so the reported row uses a maximum output length of 2,048 tokens. The five-episode pilot includes four malformed actions and therefore serves as a compatibility and 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 excluded from the main results because of its limited scope.

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 a maximum output length of 2,048 tokens. Its small size and malformed actions define its scope as a compatibility and boundary check.

Model	Success	Locally 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 within the same low-success regime: relaxing Alloy composition heavy-scaling and using the GB1 additive evaluator both raise success, yet absolute success remains low and structural events remain common. The memory sweep shows heterogeneous effects, especially on Word Ladder, while GB1 remains hard under every condition. Summary reaches 4/20 on Alloy composition, whereas best-state drops to 1/20, and both preserve the GB1 conclusion. The Alloy composition gains under randomized and misleading history provide evidence of variance and prompt perturbation. In a targeted GPT-4o sensitivity stress test over seeds 0–9, Alloy composition 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 composition setting dependence visible in GPT-4o-mini, while GB1 remains at 0/10 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

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 composition sensitivity	Default	4/50	77%	415
Alloy composition 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 composition memory	State only	0/20	78%	188
Alloy composition memory	Full history	2/20	72%	162
Alloy composition memory	Summary	4/20	71%	154
Alloy composition memory	Best-state	1/20	73%	166
Alloy composition memory	Randomized history	2/20	74%	162
Alloy composition 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

only to 5/5 full-history, but also reaches 4/5 under randomized history and 3/5 under misleading history; Alloy composition 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 perturb or improve behavior, with effects that remain insufficient for reliable 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 API stability rerun on the main environments using 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 composition	Zero shot	25%	0%	-25
Alloy composition	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 environments shows larger absolute movement for Word Ladder and Alloy composition than for GB1. The qualitative GB1 pattern is stable: local edit validity remains 100%, but success remains near zero. These results motivate reporting qualitative patterns, paired seeds, and raw logs, with fine-grained hosted model API rankings left unspecified.

Table 22: OpenAI GPT-4o-mini decoding-robustness pilot on the three main environments, 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 composition	Zero shot	5/20	66%	159
Sampled	Alloy composition	Zero shot	2/20	65%	178
Deterministic	Alloy composition	Self check	1/20	79%	174
Sampled	Alloy composition	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

This sampled-decoding pilot provides a single-model sensitivity check that motivates deterministic decoding as a protocol choice: sampled decoding leaves Word Ladder close in success but lowers zero shot local validity, reduces Alloy composition success, and leaves GB1 at 100% local edit validity but near-zero success under both decoding settings. The open family analysis below provides the corresponding full sampled-decoding robustness grid.

C.13 Open-Family Robustness Analyses

We complement the single-model pilots above with a five-experiment open family panel that uses the same three open-weight models of approximately 8B parameters (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 according to the experiment-specific evaluation budgets reported below. The panel tests whether the patterns observed in the pilots recur across open-weight model families. The released code includes the aggregation procedure and corresponding raw per episode logs needed to reproduce these results.

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-like	Llama-3.3-70B	Llama	26%
GB1	Ministral-8B	Mistral	22%

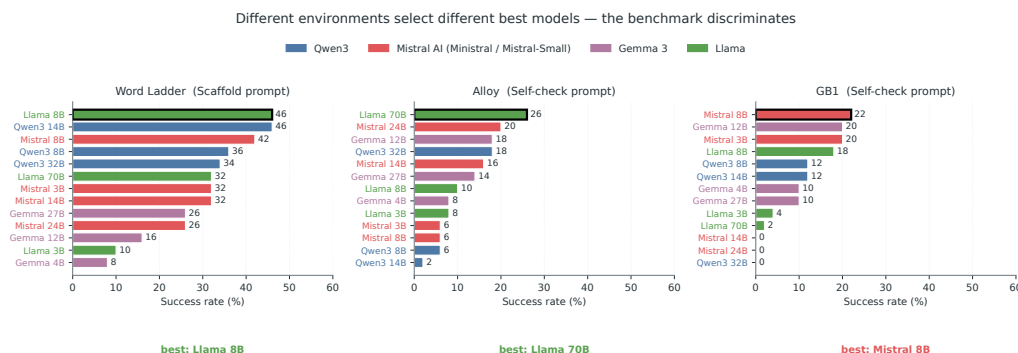


Figure 3: Different environments select different best models in the 13 model open family panel under the structural prompt. The panel characterizes environment dependent discrimination across the evaluated models.

Table 24: Three benchmark properties supported by the open-family robustness analyses. *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-like but is last on GB1	Fig. 8
Stability (no qualitative flip)	Scorer ablation: cell-level Δ success up to ≈ 25 pp, 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, environment) 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

Scorer ablation. Figure 4 and Table 25 report the same shallow wrapper grid as Table 15 but varies the scorer among a visible state scorer, a proxy-distance scorer, and a privileged oracle scorer. The first two scorers use visible state information; the oracle scorer uses privileged information as a diagnostic sensitivity condition. Its effects are nonmonotonic: Llama Word Ladder drops from 60% under visible state scoring to 45% under privileged scoring. This supports the interpretation of scorer choice as a sensitivity analysis. Llama Alloy composition 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 composition proxy.

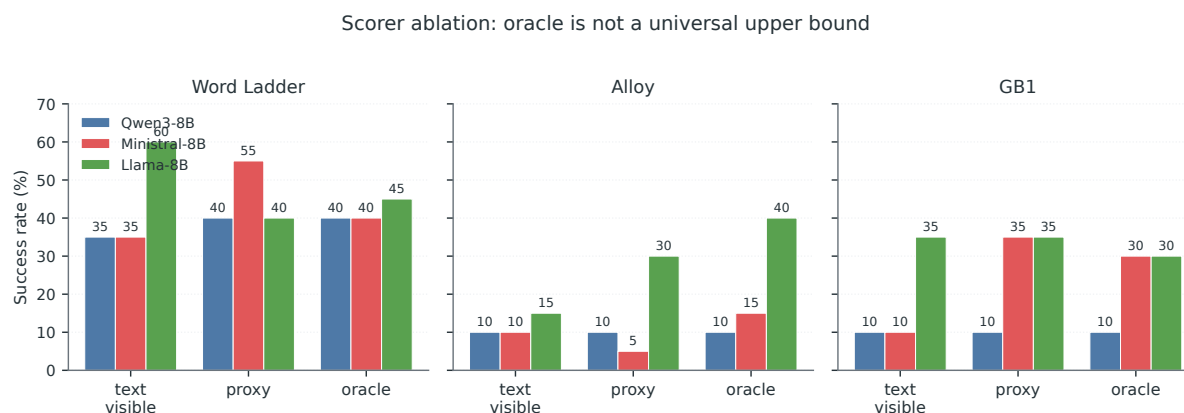


Figure 4: Open-family scorer ablation. Each panel shows visible state, proxy-distance, and privileged scorers over 20 seeds for the three approximately 8B models. Privileged target-distance scoring has nonmonotonic effects: Llama Word Ladder drops from 60% under visible state scoring to 45% under the privileged scorer.

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 a bounded three-step window (54%) and is hurt by full history (30%). Alloy composition 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), locating the primary bottleneck beyond memory access. Across environments, the best memory depends on both model and environment, which is why the main paper treats state only memory as the conservative reporting choice.

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

Model	Environment	Visible state	Proxy	Privileged	Calls/episode
Qwen3-8B	Word Ladder	35%	40%	40%	50.7
Qwen3-8B	Alloy-like	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-like	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-like	15%	30%	40%	24.0
Llama-8B	GB1	35%	35%	30%	16.5

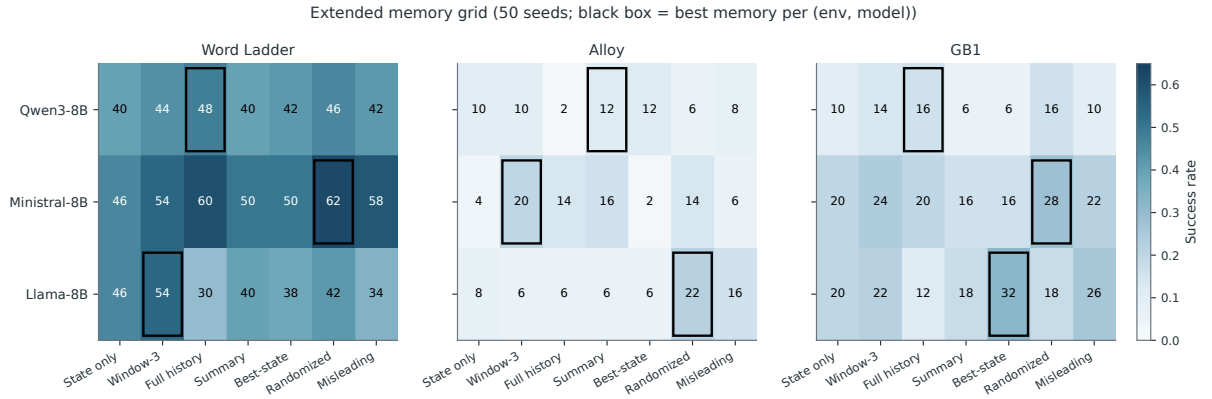


Figure 5: Open-family extended-memory grid (50 seeds, structural prompt, deterministic decoding). Black boxes mark the best memory module per (environment, model). The best memory differs across every row in every panel: full history is best for Qwen3 on Word Ladder, a three-step window for Llama, and randomized history for Ministral; the alignment changes again on Alloy composition and GB1.

Table 26: Open-family extended-memory grid. Each cell is success rate over 50 seeds with the structural prompt and deterministic decoding. The randomized-history condition shuffles the trajectory history before each call; the misleading-history condition 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-like	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%

Surrogate sensitivity. Figure 6 and Table 27 report the open family version of the GPT-4o-mini sensitivity sweep. The Alloy composition heavy-scaling toggle raises mean success modestly (Qwen3 zero shot from 2% to 12%, Llama zero shot from 0% to 10%) while preserving 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%). The consistent GB1 difficulty across evaluators strengthens the landscape result. Both directions show that surrogate settings perturb levels while preserving the qualitative structure.

Surrogate sensitivity: setting toggles do not flip qualitative claims

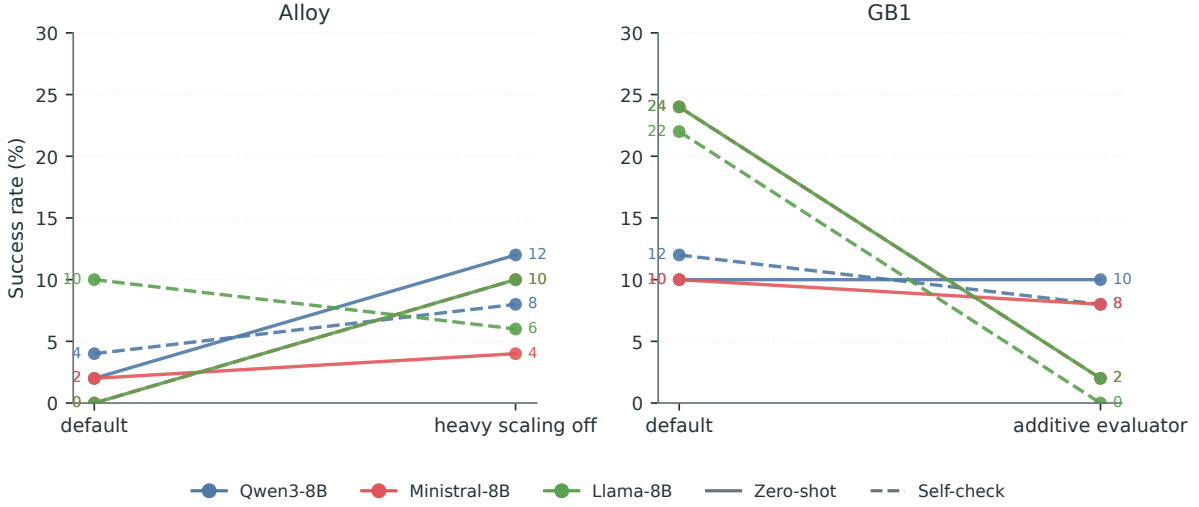


Figure 6: Open-family surrogate-sensitivity slope chart. Lines connect each (model, prompt) cell’s success rate under the default and perturbed settings. The Alloy composition heavy-scaling toggle lifts most cells; the GB1 additive evaluator drops most cells. Both directions stay inside the low-success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

Table 27: Open-family surrogate sensitivity on Alloy-like (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-like	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%

Hosted model API stability. Table 28 extends the date-separated API stability check from the document plan pilot (Table 20) to the main-environment open family panel. Cell-level absolute success varies by up to 25 percentage points relative to other deterministic Alloy composition and GB1 runs in this appendix, motivating the main paper’s emphasis on qualitative patterns and raw logs for hosted APIs.

Table 28: Open-family hosted-model API stability rerun on Alloy-like 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-like	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 preserves the central qualitative patterns: Word Ladder remains the highest-success environment, Alloy composition 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, motivating paired-seed reporting and matched protocol comparisons in the main paper.

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-like	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 5.3 and 5.5. Across these conditions, format only few shot prompting yields inconsistent gains, while larger gains appear when the intervention changes the editing policy in an environment specific way.

Table 30: Detailed main prompt intervention success rates (%) under state only history ($M0$), with Wilson 95% confidence intervals in brackets.

Environment	Zero shot	Few shot (format)	Few shot (strategy)	Scaffold	Self check (generic)	Self check (task-specific)
Word Ladder (four letter)	18 [10,31]	8 [3,19]	18 [10,31]	30 [19,44]	6 [2,16]	18 [10,31]
Alloy composition	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 composition 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 composition, and 1,866 versus 321 in GB1. This supports the diagnostic claim that the benchmark exposes delayed trajectory construction failures beyond action formatting. The cross model table also shows heterogeneous best prompts and size effects across model families in these constrained editing environments.

The released resources also include 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, conditioned on surface cleanliness success, and trajectory diagnostics. The intentionally unbalanced model grid supports qualitative recurrence; estimating a scaling law or leaderboard would require balanced family coverage.

Table 31: Cross model success rates (%) under zero shot / format only / structural prompting. Absolute performance varies across families and scales. Format only prompting yields inconsistent gains, while delayed structural events dominate Alloy composition and GB1 failures.

Model	Word Ladder	Alloy composition	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, environment, 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. The best row varies by environment, further showing heterogeneous cross family scaling on the scientific-design environments.

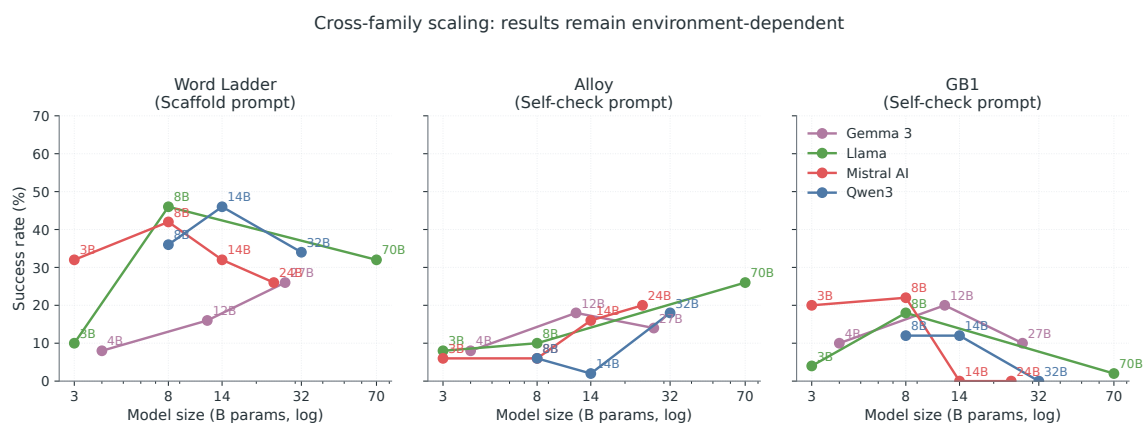


Figure 7: Cross-family scaling curves. Each panel plots success rate against model size (log scale) for one environment under the structural prompt. Line markers annotate each model’s parameter count. Mistral / Ministral are merged into a single developer-family line because they share a developer but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

F Detailed History Ablation

This section reports the full history ablation results summarized in Section 5.6. Recent edit history has conditional effects: it helps when the task rewards cumulative state tracking and harms performance when recent trajectories reinforce a poor local path.

Table 32: History-ablation success rates (%) for selected prompt conditions.

Environment	Prompt	<i>M0</i>	<i>M1</i>	<i>M3</i>	Interpretation
Word Ladder	zero shot	22	20	10	harmful
Word Ladder	scaffold	30	18	10	strongly harmful
Alloy composition	zero shot	12	20	25	helpful
Alloy composition	self check	18	8	18	nonmonotonic
GB1	zero shot	25	8	8	strongly harmful
GB1	self check	35	5	5	catastrophically harmful

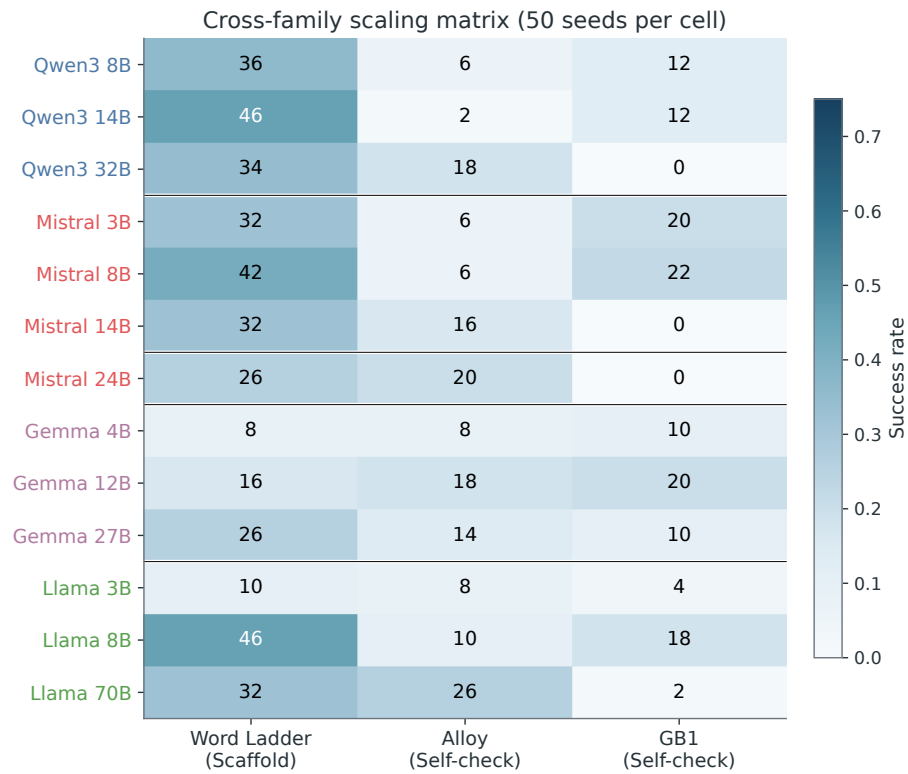


Figure 8: Cross-family scaling matrix: 13 models grouped by family and size against the three environments under the structural prompt, 50 seeds per cell. The y-axis labels are color-coded by family. The leading row varies by environment: Llama 70B leads on Alloy composition, Mistral 8B leads on GB1, and Qwen3 14B ties Llama 8B at the top of Word Ladder.

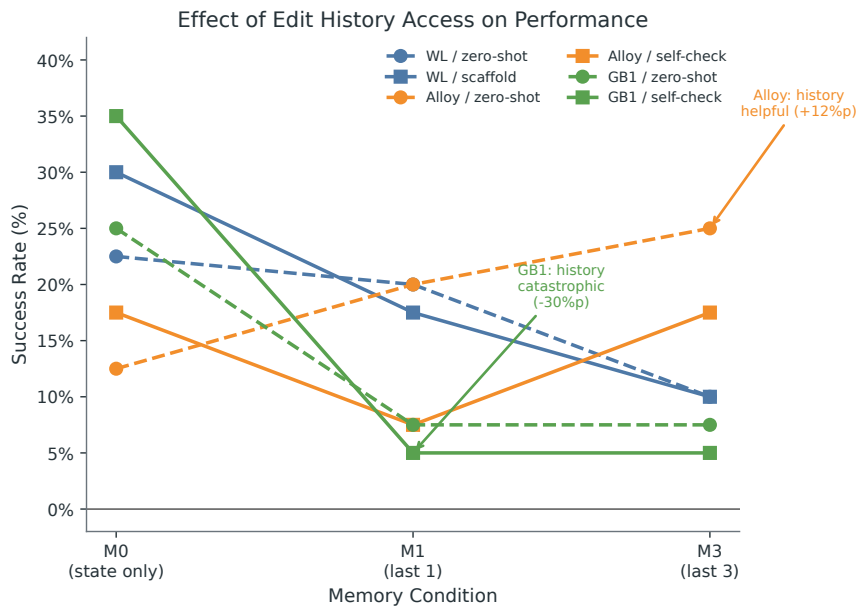


Figure 9: History ablation for selected prompt conditions. Recent trajectory access helps only when the raw history carries useful cumulative signal; it can also reinforce poor local commitments.

G Failure Patterns

This section gives the dominant failure patterns underlying the discussion in Section 5.2. The main residual errors vary by environment and are predominantly structural.

Table 33: Representative dominant failures under the main prompt experiment. Residual errors are predominantly structural, with a different dominant mode in each environment: budget-unaware continuation in Word Ladder, objective-tradeoff collapse in Alloy 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 composition	zero shot	OBJECTIVE TRADEOFF (20)
Alloy composition	few shot (format)	OBJECTIVE TRADEOFF (21)
Alloy composition	few shot (strategy)	OBJECTIVE TRADEOFF (23)
Alloy composition	scaffold	OBJECTIVE TRADEOFF (22)
Alloy composition	self check (generic)	OBJECTIVE TRADEOFF (20)
Alloy composition	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 guidance content.

H.1 Lexical Editing: Word Ladder

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 Composition Editing: Alloy composition Proxy

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 $\geq 2,000$)
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 Protein Sequence Editing: GB1 Landscape

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 composition Exemplar Robustness

This section reports the Alloy composition exemplar-sensitivity analysis referenced in the main text. Prompt effects in Alloy composition editing depend on both the presence of a demonstration and the specific trajectory it induces.

Table 34: Alloy composition few shot exemplar robustness.

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

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 nontrivial but non-degenerate regime.

Table 35: GB1 operating point calibration across fitness thresholds. Values are success rates (%). Additive local heuristics fail across all tested thresholds, whereas the LLM and the privileged noisy policy remain above zero. This separates the chosen regime from additive local improvement and confirms attainable success under stronger policies.

Policy	Threshold 4.7	Threshold 4.8	Threshold 5.0
Additive greedy	0	0	0
Observation-level greedy	0	0	0
Zero shot LLM	30	45	40
Few shot strategy	20	20	20
Privileged noisy policy	55	55	55

K Word-Length Ablation

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

Table 36: Word-length ablation for lexical editing. The four 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
four letter	13	27	+14
5-letter	10	13	+3