CANDIDATE ADMISSION GATES FOR GENERATIVE REASONING: DIAGNOSTICS AND VMGR PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Which change does a performance gain in generative reasoning identify? We study this question through comparison diagnostics and a controlled generative-planning method. Three formative comparisons of autoregressive and iterative systems expose distinct population, competence, measurement and provenance failures in code repair, GF(2) reasoning and minimal planning. We organize these failures into six candidate admission gates for bounded, verifiable tasks. A claim-specific state representation and CPU replayer distinguish blocked attribution from reportable system outcomes; the cases motivate, but do not independently validate, the procedure. A later planning diagnosis finds that selecting preferred demonstrations also changes candidate coverage. VMGR separates the two: verified trajectories define shared support, budget-conditioned completion costs define soft targets, and a uniform floor retains every candidate in a root-balanced loss. On Blocksworld with two-action macros, VALUE exceeds same-support UNIFORM by 3.646 percentage points on 768 new exact task identities, using one 0.5B language model and three paired rank-64 fits. Later search recovery reaches 99.94% on 600 smaller public tasks, but almost removes the weighting gap. On all 110 tasks in a harder public configuration, frozen VALUE and same-search UNIFORM reach 62.73% and 52.12%. The 10.606-point difference has a conditional task-cluster 95% interval of [4.848, 16.667], alongside 14.34% less inference token-work. Thus the strongest method evidence is a same-support supervision gain, not the larger gap against poorly transferring SFT controls. The study supports a scoped planning improvement and useful comparison distinctions, not universal gate validity or architecture-level superiority.

1 INTRODUCTION

An accuracy gap is an observation, not yet an explanation. In generative reasoning, it can reflect the task population, representation, verifier, training exposure, decoder or implementation, as well as the model. We call their combination a **system package**. The research question is therefore not only which package wins, but which change the comparison identifies. A useful study should preserve both answers without confusing them.

Our initial autoregressive (AR) versus iterative comparisons found an unavailable population, a shared exact-success floor, and invalid measurement combined with a recipe mismatch. Six candidate admission gates distinguish these blocked claims from valid but limited package results. These diagnoses require different repairs; they are not six new principles of validity.

A later planning study turns this diagnostic question into a controlled intervention. Value-aligned and lexical demonstration constructors produced non-nested supports of 5,028 and 5,592 root–prompt pairs. Preference and coverage changed together. This motivated a narrower question: **can completion costs improve a generative policy when candidate support is held fixed?**

VMGR forms a replay-verified union of complete trajectories, computes budget-feasible continuation costs and converts them into training preferences. A uniform floor retains every candidate in a root-balanced generative loss. The UNIFORM control uses the same support, updates and decoder, changing only the target weights. Both policies generate new macro text at inference; neither retrieves plans from the training graph.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Two contributions connect diagnosis to intervention: a claim-specific blocking and reporting procedure, and verified shared-support supervision with a paired generative-planning test. Experiments separate weighting, search recovery and package transfer. VMGR makes this supervision contrast interpretable; it neither repairs the original cases nor introduces a new architecture or weighting principle. Success is primary; resources locate each operating point.

2 From comparison failures to a constructive study

2.1 Candidate conditions and dependency-aware decisions

An admission gate records whether the available evidence supports a **declared claim**. It neither requires a positive effect nor certifies validity as a binary property. The six conditions in Table 1 were synthesized retrospectively, not selected through independent validation.

Gate	Required distinction	Application to the later VMGR study
G1: population	Intended population versus available tasks	Separate fitting, reused development and exact-identity-new tasks.
G2: semantics	Valid objects and verifiers versus proxy scores	Replay executable plans; disclose macro versus primitive interfaces.
G3: competence	Meaningful contrast versus a shared floor	Keep failed adaptations diagnostic, not evidence against entire methods.
G4: resources	Declared package budgets versus equal-work attribution	Match VALUE–UNIFORM training rules; disclose unequal SFT budgets.
G5: measurement	Intended construct versus an invalid statistic	Measure verified goals and preserve paired task clusters.
G6: provenance	Claimed recipe versus executed implementation	Separate published, released and adapted recipes and fixed endpoints.

Table 1. Candidate comparison conditions and design responses, not a pass certificate. Their relevance depends on the contrast. Unequal resources can support a declared system tradeoff, but not an equal-resource claim.

The state representation distinguishes passed, failed, invalid, unreached, unavailable and non-applicable prerequisites. Failed or unavailable applicable prerequisites block dependent claims; a local pass alone cannot admit a claim. The CPU artifact checks supplied labels and replays three stopping rules. Six synthetic records and 16 tests exercise seven states and five terminal labels (Appendix F), not scientific validity or false-stop rates.

For example, same-support VALUE–UNIFORM isolates weighting for fixed fits. Hard-110 versus differently trained SFT ranks executed packages, but cannot identify weighting or refute SFT in general. The same records permit one statement while blocking another, rather than assigning a global verdict to a study. Appendix F.1 works through both claims.

2.2 Three formative diagnostic cases

Code repair: population and interface limits. TRAIN engineering qualification succeeded, but a frozen source census supplied 870 eligible rows for 3,712 required candidate positions. Bounded text interfaces also failed qualification. This stopped the intended fresh-problem comparison; it did not rank generative architectures.

GF(2): a shared competence floor. Matched 170M packages obtained zero exact passes on 1,440 paired formal systems. New-seed controls did not establish prerequisite task-and-interface competence. A zero difference at this floor does not establish equivalence.

Minimal planning: measurement and provenance. Unchanged-output replay corrected a reversed difficulty stratifier. The corrected easy AR condition passed, while the iterative condition remained at zero exact and semantic pass. The audited public script specifies 100 epochs, whereas the paper specifies 1,200 (Ye et al., 2025; HKUNLP, 2025). Our local run produced 1,000 unsuccessful

outputs. The invalid record and corrected successor remain separate. Failure of the executed recipe does not refute the different published recipe.

None passed the complete procedure. The later planning diagnosis is distinct: preference and coverage changed together. Fixing their union makes preference testable. This motivates VMGR, without validating the gates or repairing the three original cases.

3 VMGR: CONTINUATION COSTS WITHOUT SUPPORT FILTERING

Figure 1 separates shared support, target weighting and deployment.

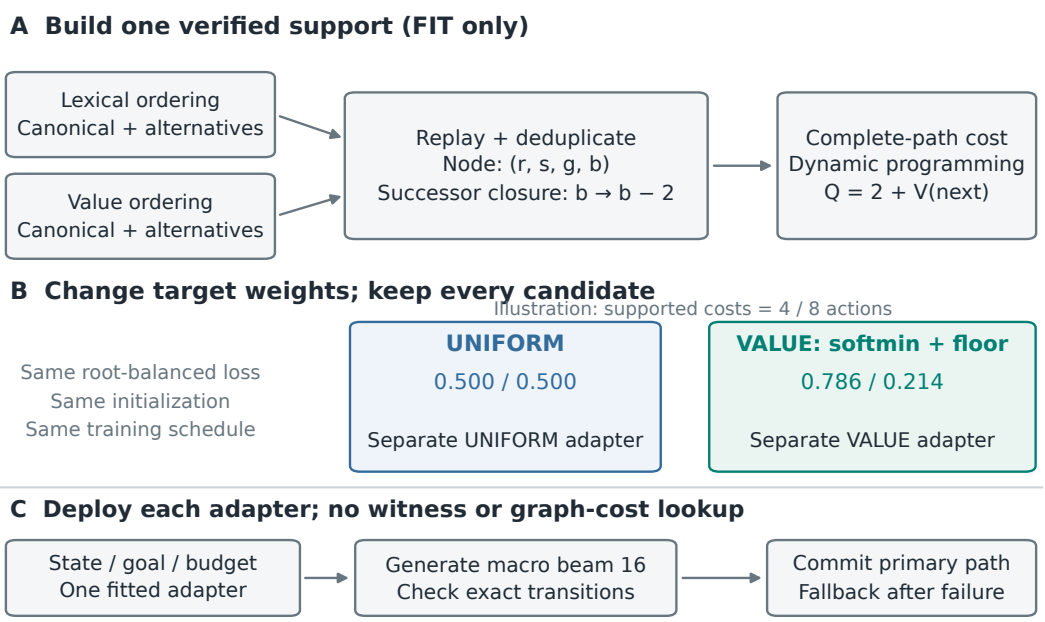

Figure 1: **Separate coverage, preference and execution.** Two source constructors provide complete witnesses. Their replay-checked union defines one budget-conditioned graph. VALUE and UNIFORM use the same candidates, root-balanced loss reduction and schedule to fit separate adapters. Only target weights differ. The four/eight-action example is illustrative. Deployment uses generated macros and exact transitions, not stored witnesses or graph-cost lookup. The later fallback uses retained siblings after primary failure.

3.1 FROM COMPLETE WITNESSES TO A SHARED GRAPH

A training root r defines an initial Blocksworld state and goal g. A prompt $x = (s, g, b)$ encodes the current state s and remaining primitive-action budget b. Each candidate macro a contains two primitive actions and ends with an empty hand. The graph identifies a node by (r, x), so different roots or remaining budgets are never silently merged.

Two constructors supply replayed completions from 72 fitting roots. They rank first macros lexically or by witnessed completion length, retaining a canonical witness and up to three alternatives. VMGR takes their deduplicated union and exposes every two-action transition (Appendix A.2).

Edges must execute legally and reach either the goal or an existing successor at budget $b - 2$ under the same root. Missing successors and budget violations invalidate construction; no search fills gaps. Decreasing budgets make the graph acyclic even when a physical state recurs.

Let V_G be the shortest remaining completion in the observed graph. For a candidate with successor x',

$$Q_G(r, x, a) = 2 + V_G(r, x'), \qquad V_G(r, x) = \min_{a \in A(r,x)} Q_G(r, x, a). \tag{1}$$

A terminal goal node has $V_G = 0$, so a goal-reaching two-action macro has $Q_G = 2$. Dynamic programming evaluates nodes in increasing remaining-budget order. The result measures a complete supported path, not a one-step heuristic. It is an upper bound on unrestricted optimal completion cost, since missing edges may contain shorter solutions. This restriction is central: exact replay certifies feasibility, not exhaustive coverage.

3.2 Cost-sensitive targets with a uniform floor

For $K = |A(r, x)|$ candidates, UNIFORM assigns $w_U(a \mid x) = 1/K$. Within a root, let $Q_{\min}(x) = \min_{a \in A(r,x)} Q_G(r, x, a)$. We suppress the root index in the target formula below. VALUE assigns

$$w_V(a \mid x) = \frac{\epsilon}{K} + (1 - \epsilon) \frac{\exp[-(Q_G(x, a) - Q_{\min}(x))/\tau]}{\sum_{a' \in A(r,x)} \exp[-(Q_G(x, a') - Q_{\min}(x))/\tau]}. \tag{2}$$

We use $\epsilon = 0.25$ and $\tau = 2$ primitive actions, fixed before the shared-support model evaluations. The minimum-cost baseline improves numerical stability and cancels under normalization. Tied costs produce uniform targets.

The floor guarantees at least ϵ/K objective mass, independently of cost gaps. It does not bound learned probabilities or beam diversity. Softmin alone is already positive in exact arithmetic. Figure 1 illustrates the weights for costs four and eight; Appendix A.1 gives the limits and baseline invariance.

The corpus has 6,657 nodes and 9,428 edges. **VALUE versus UNIFORM** changes targets at 834 unequal-cost nodes, or 11.643% of root-balanced objective mass. Most prompts remain identical. A different, post-hoc comparison asks whether graph costs add information beyond the shortest recorded witness suffix. **Graph cost versus witness cost** changes 357 soft targets but only one minimum-cost candidate set (Appendix A.5). These counts do not describe VALUE versus UNIFORM. That diagnostic measures additional graph information; no trained-model ablation isolates its contribution.

3.3 Root-balanced generative learning

Let R be the training roots and X_r the distinct prompts of root r. Let $\ell_\theta(a \mid x)$ be mean suffix-token negative log likelihood, including supervised EOS and excluding prompt and padding tokens. Both arms minimize

$$\mathcal{L}_w(\theta) = \frac{1}{|R|} \sum_{r \in R} \frac{1}{|X_r|} \sum_{x \in X_r} \sum_{a \in A(r,x)} w(a \mid x)\, \ell_\theta(a \mid x). \tag{3}$$

Equal-root reduction prevents larger recorded supports from dominating. Each update covers nine roots, accumulating every candidate in microbatches of at most four. Global coefficients remain unchanged; there is no local renormalization or target pruning. All targets have 15 supervised tokens, including EOS.

3.4 Generative execution and optional fallback

Both policies use ordinary beam 16, at most 16 new tokens per call and repetition penalty 1.1. After parsing, execution, deduplication and path-local cycle filtering, the decoder commits the highest-scoring admissible macro. It stops at the goal, a failed candidate set or 36 primitive actions.

The later fallback preserves this primary policy and retains unused legal siblings. After failure it accepts a retained goal or expands the frontier. Priority sums native length-normalized beam scores, not path log probabilities. Each branch uses its remaining action budget. There is no learned critic, distance oracle or reference-plan access. Public-600 permits 36 total calls and 36 actions per branch. Hard-110 permits 128 calls and 64 actions for all three compared systems, with fixed fits. Fallback is evaluated separately from the original decoder.

4 EXPERIMENTAL DESIGN

4.1 FIXED MODELS, DATASETS AND EVALUATION CHRONOLOGY

We use Qwen2.5-0.5B-Instruct (Qwen et al., 2024) with LoRA (Hu et al., 2022) rank 64, alpha 128, zero dropout and 35,192,832 trainable parameters. The backbone is FP16; adapters and suffix cross-entropy are FP32. Each shared-support fit receives 22 epochs and 176 AdamW updates: 2×10^{-4} learning rate for updates 1–160 and 2×10^{-5} for 161–176, retaining optimizer moments. VALUE and UNIFORM share three initialization seeds, root schedules, support and prescribed work. Panels reuse these paired fits (Appendix A).

Table 2 separates fitting, recipe development and evaluation. Rank 64 was selected after development studies of adapter ranks 16, 32, 64 and 128. Those studies are not fresh confirmation or backbone scaling.

Population	Tasks	Blocks	Role and chronology
FIT	72	6–11	Witness construction and optimization
TRAIN-selection	72	6–11	Separate roots; repeated recipe development
New-root panel	768	6–11	Models, decoder and analysis fixed before outcomes
Public-600	600	3–5	Initial comparison; later controls and fallback use observed tasks
Hard-110	110	6–15	Systems frozen after Public-600; no further fitting

Table 2. Data roles, not interchangeable replications. All evaluation tasks are paired across three fits. New-root and original Hard-110 settings were locally frozen before their outcomes. The native-feedback extension in Appendix D.5 uses already observed Hard tasks; it is not fresh confirmation.

The new-root panel retains the first 128 eligible, initially unsolved roots per block from 6,144 generated candidates. Exact identities in registered accessible history are excluded without using model outcomes or oracle difficulty (Appendix A.3).

Public-600 contains all 500 generated-basic and 100 generated-basic-3 tasks at a pinned PlanBench revision (Valmeekam et al., 2023). It has no exact overlap with 185,416 accessible historical identities. Hard-110 is the complete source hard configuration and has no exact overlap with Public-600 or 2,682 fitting prompt states. Other historical exclusions are not certified for Hard-110. Exact identity checks establish neither non-isomorphism nor pretraining decontamination (Appendix D.4).

4.2 CONTROLS AND MEASUREMENT

UNIFORM is the identifying weighting control. A secondary PERMUTED arm keeps each node's candidates and weight multiset but permutes the assignment of weights to candidates. One map, frozen with seed 20260923, changes 628 nodes and 9.160% of root-balanced objective mass. Three new fits use the same 176-update recipe. This post-hoc mechanism study reuses the observed 768-root panel and averages over neither new task populations nor permutation maps.

The public SFT controls share the backbone and adapter size but have different training targets and exposure. Stepwise and complete-plan SFT each receive a fixed target of four million active input tokens per fit. Their checkpoints are not retuned for these comparisons. We evaluate greedy/native-feedback, sampled Best-of-16, receding-horizon verified-step and best-first decoding. Best-first SFT uses beam 16, up to 36 expansions and 36 primitive actions. It is an adapted search control, not an author-faithful ToT or RAP reproduction. A later Hard-110 native-feedback control reuses complete-plan fits, with 15 rounds and a 64-action cap. Appendices B–D specify settings, feedback and costs.

Success requires legal execution and terminal goal satisfaction. Each task cluster contains all three paired fits. We use 20,000 bootstrap draws: new-root estimates weight block strata equally and resample within strata; public estimates weight and resample tasks equally. Percentile 95% intervals condition on the fixed fits. Appendix A.3 gives seeds and replay scope.

New-root effects were read only after all six cells completed. Missing records abort analysis rather than count as task failures. There was no effect-size gate, seed deletion or optional stopping. Per-fit,

leave-one-fit-out and crossed root/seed analyses expose sensitivity. Equal compute is not required to report success; unequal resources still require disclosure.

5 Results: Preference Gains and Recoverable Search Failures

5.1 Cost Weighting Improves the Fixed-Decoder Mean

On new roots, VALUE solves 1,957/2,304 paired positions (84.939%) versus 1,873 (81.293%) for UNIFORM. The gain is **3.646 percentage points**, with a conditional 95% interval of [2.040, 5.295]. VALUE wins 220 paired positions and loses 136. All six block-stratum point differences are positive. Figure 2 retains the variation hidden by that mean. Table E1 gives every fit's success counts and differences.

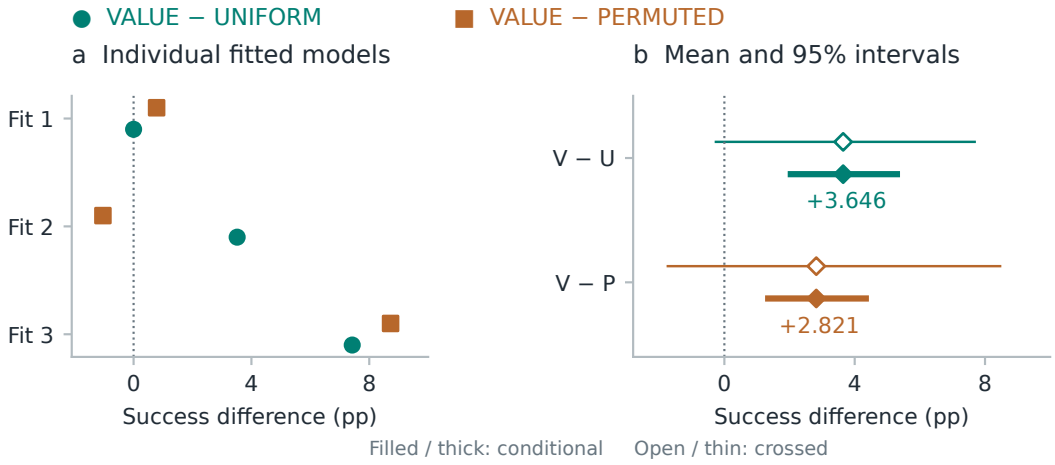

Figure 2: **Conditional gains and seed sensitivity answer different questions.** Left: fixed-fit VALUE minus UNIFORM and VALUE minus PERMUTED effects, in percentage points, on the same 768-root panel. Right: pooled estimates with conditional task-cluster and crossed root/seed 95% intervals. The primary comparison was fixed before new-root outcomes; permutation is a later single-map mechanism probe. Crossed intervals include zero for both contrasts.

Leaving out each fit gives VALUE–UNIFORM differences of +5.469, +3.711 and +1.758 points. The crossed root/seed sensitivity interval is [−0.260, 7.682]. The mean gain is supported conditional on these fits; improvement over future training seeds remains unresolved. An earlier rank-16, beam-4 study also had a positive pooled contrast, +1.953 points [0.303, 3.602], but mixed fit signs. Table E3 retains both studies without pooling populations or attributing their difference solely to capacity.

VALUE exceeds PERMUTED by 2.821 points [1.345, 4.340]. However, its per-fit differences are +0.781, −1.042 and +8.724. Omitting the strongest fit reverses the mean to −0.130, and crossed sensitivity spans [−1.736, 8.464]. UNIFORM–PERMUTED is −0.825 points [−2.257, 0.608]. This control gives seed-dependent evidence consistent with useful cost alignment, not proof that alignment explains every gain or that the floor is necessary.

5.2 The Complete Public System Reaches 99.94%

VALUE with fallback solves **1,799/1,800 executions (99.944%)**, versus 1,740 (96.667%) for best-first SFT. The gain is **3.278 points [2.222, 4.444]**, with 60 paired wins and one loss. All three fit differences and leave-one-fit-out means are positive. Table 3 leads with this complete system, then separates its no-fallback ablation from other operating points. Fallback was developed on the observed panel; this is a system-development result, not an independent confirmation.

The stronger control matters: best-first raises the same stepwise SFT fits from 948 to 1,740 successes. Its 44.000-point gain also changes terminal-child validation, not only backtracking. Native

Modulo raises its own Direct result from 958 to 1,003, using additional feedback rounds and computation.

Public operating point	Successes /1,800	Success, %	Role
VALUE + fallback	1,799	**99.944**	Later search extension
UNIFORM + fallback	1,794	99.667	Same-extension control
Stepwise SFT + best-first	1,740	96.667	Strong adapted search
VALUE, original macro beam	1,715	95.278	Fixed-decoder method
UNIFORM, original macro beam	1,669	92.722	Matched weighting control
Complete-plan SFT + Best-of-16	1,469	81.611	Verified sampling
Complete-plan SFT + native Modulo	1,003	55.722	Fitted feedback adaptation
Stepwise SFT + verified-step beam	948	52.667	Receding-horizon control

Table 3. Verified public success at declared operating points. All rows use the same 600 tasks and three fixed fits per policy. Only VALUE–UNIFORM within a decoder matches support, training work and inference rules. System rows differ in exposure, representation and compute. The two fallback rows are post-hoc development on the observed panel.

5.3 Search and weighting make different contributions

Without fallback, VALUE–best-first is −1.389 points [−3.000, 0.167]: the interval supports neither a directional difference nor equivalence. This ablation does not replace the complete-system result in Section 5.2. VALUE–UNIFORM is +2.556 points [1.389, 3.778] (Table 3); Appendices C and D.2 retain the paired breakdown.

Fallback reuses all 3,384 historical successes and reconstructs all 216 failures, retaining the 3,600-position denominator. Reconstructed primaries match recorded actions, termination and counts. New branches and duplicated prefixes are charged (Appendix D.3).

It recovers **84/85 VALUE failures** and **125/131 UNIFORM failures** (Figure 3). Same-fallback UNIFORM reaches 1,794/1,800. The residual weighting gain is 0.278 points [0.000, 0.556], with six wins and one loss; omitting the second fit removes it. Search recovers headroom for both objectives, so the final system lead cannot be attributed to weighting alone.

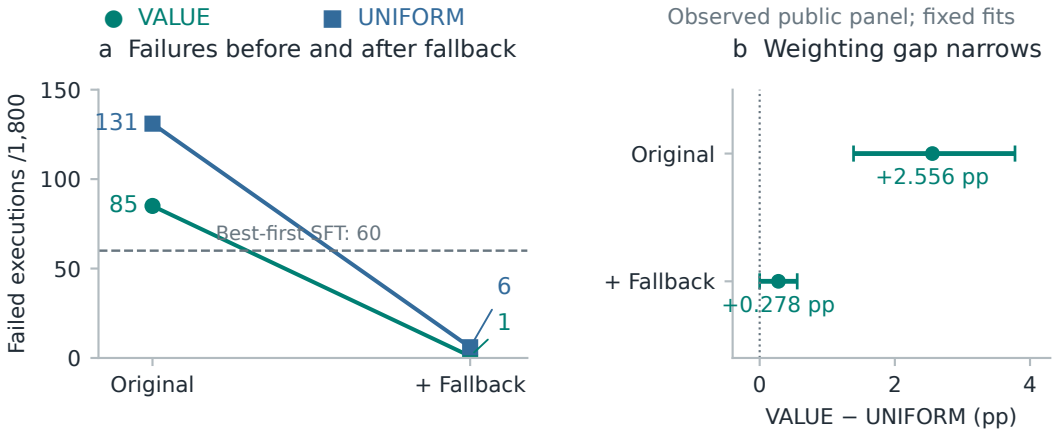

Figure 3: **Search recovers most remaining public failures.** Left: exact unsuccessful execution counts before and after fallback. The dashed line shows the 60 failures of best-first SFT. Right: VALUE–UNIFORM success differences with conditional task-cluster 95% intervals. All results use 600 tasks and three fixed fits. Original denotes the no-fallback ablation. The extension reuses old successes, so zero regressions are structural. The narrowing contrast is descriptive, not a separately tested causal interaction between training weights and search.

5.4 HARDER TASKS RETAIN A WEIGHTING ADVANTAGE

On Hard-110, VALUE with fallback solves **207/330 executions (62.73%)**, versus 172 (52.12%) for UNIFORM and 13 (3.94%) for best-first SFT. The same-search VALUE–UNIFORM contrast is **+10.606 points [4.848, 16.667]**, with 68 wins and 33 losses. This is the identifying supervision comparison. It remains the protocol's secondary endpoint: the frozen primary system contrast, VALUE–SFT, is +58.788 points [52.121, 65.455], with 194 wins and no losses. We do not relabel endpoints after observing them, or attribute the larger system gap to cost weighting. The analysis resamples 110 task clusters, not 990 independent tasks.

Hard-110 fit	VALUE /110	UNIFORM /110	SFT /110	V−U, pp	V−SFT, pp
20261633	72	64	1	+7.273	+64.545
20261634	61	49	4	+10.909	+51.818
20261635	74	59	8	+13.636	+60.000
Total /330	207	172	13	+10.606	+58.788

Table 4. Frozen-fit transfer to the complete harder public configuration. VALUE and UNIFORM share support, training exposure and search rules. SFT is the fixed adapted best-first control; it has different training and action representation. No endpoint is selected on Hard-110 outcomes.

All leave-one-fit-out weighting means remain positive, from +9.091 to +12.273 points. This does not imply uniform benefits: the 12-block stratum favors UNIFORM by 5.128 points. All unsuccessful executions exhaust the expansion budget: 123 for VALUE, 158 for UNIFORM and 317 for SFT. The control that was strong on the smaller panel transfers poorly here. Thus Hard-110 shows poor transfer for these SFT fits, not failure of SFT or tree search in general. The larger weighting gap than on Public-600 is descriptive: both population and budgets change between panels.

A follow-up native full-plan feedback control solves **4/330**, all on the first attempt. VALUE leads by +61.515 points [54.848, 68.182], with 203 wins and no losses. This observed-panel extension shows poor transfer of the retained fits, not superiority over a competent external-SOTA reference. Appendix D.5 retains all strata, failures and costs.

5.5 HIGHER SUCCESS NEED NOT REQUIRE MORE INFERENCE WORK

On Hard-110, VALUE's 10.606-point advantage over same-search UNIFORM accompanies 13.851% fewer generation calls and 14.340% fewer inference token-work positions (Table E2). Both policies process 101,415,996 training input tokens across three fits. The gain therefore does not require greater measured inference work at this matched-rule operating point; it does not establish a full cost frontier.

Against best-first SFT, VALUE uses 73.497% less Hard-110 inference token-work, but training costs 21.020 versus 4.421 worker-hours. Native feedback uses 5.889 million inference positions and solves four executions. Higher success can justify resource tradeoffs; it does not erase them. On new roots, VALUE's 3.829% token-work reduction leaves wall time essentially unchanged. Appendix A.4 retains all failure counts. Token-work is not FLOPs, latency or total research cost.

6 RELATED WORK

Measurement and comparison validity. Construct validity concerns how well a measurement represents its intended construct, not a universal yes/no certificate (Jacobs & Wallach, 2021). Comparable held-out performance can leave application-relevant behavior underdetermined (D'Amour et al., 2022). Aggregate multi-task rankings can be sensitive to benchmark changes (Zhang & Hardt, 2024). Our gates organize these familiar concerns into claim-specific blocking and reporting decisions, including claimed-to-executed recipe provenance. The cited works do not establish this six-gate procedure.

Planning-derived supervision. AWR uses advantage-weighted policy regression; QWR estimates advantages through learned Q-values (Peng et al., 2019; Kozakowski et al., 2022). RAML uses

exponentiated task rewards for structured prediction (Norouzi et al., 2016). BOOM aligns a policy with an online planner through learned values (Zhan et al., 2025). VMGR contributes verified completion targets on fixed support, not exponential weighting itself. Unlike AggreVaTe's expert cost-to-go supervision and DAgger's learner-induced data aggregation, VMGR uses one observed graph and leaves sequential distribution shift unresolved (Ross & Bagnell, 2014; Ross et al., 2011).

Symbolic guidance and recovery. ASNets learn generalized planning policies from small instances for larger same-domain problems (Toyer et al., 2018). VMGR instead studies controlled supervision for generative text policies. SymPlan combines symbolic filtering, reinforcement-learning rewards and length control for embodied planning (Shi et al., 2026). AHAT learns grounded subgoals for a symbolic planner (Liu et al., 2026). RePoT repairs generated programs from verified failure checkpoints (Mazaheri, 2026). Different populations and interfaces prevent ranking their published scores alongside ours.

Inference-time search. ToT and RAP explore reasoning branches using their own proposal, evaluation and world-model interfaces (Yao et al., 2023; Hao et al., 2023). TS-LLM uses learned values and search for decoding and training (Wan et al., 2024). VMGR obtains offline targets by graph dynamic programming, without a critic. LLM-Modulo combines generated proposals with external verifier feedback (Kambhampati et al., 2024). Our controls are adaptations, not complete author-recipe reproductions. Appendix E retains earlier RAP and unfitted-feedback diagnostics without promoting their weak operating points to SOTA references.

7 Discussion and Limitations

The central finding is that completion-cost supervision can improve a generative policy without removing candidates or increasing training exposure. Search largely removes the weighting gap on smaller tasks; Hard-110 retains it with less inference work. Supervision and recoverable decoder failures therefore explain different parts of the tested system's performance.

The method is deliberately specific. We test one 0.5B backbone, one domain and two-action macros. We have not measured floor or temperature sensitivity, perturbed support, or longer-macro generalization. Graph costs describe observed completions: an unseen shorter path can reverse their ranking. The graph also omits the decoder's history-dependent cycle filter. Relative to shortest recorded witness suffixes, graph recombination mostly refines soft weights: only one FIT node changes its minimum-cost set. That restriction concerns the incremental graph contribution, not the 834-node VALUE–UNIFORM intervention. No graph-versus-witness model ablation shows that recombination is necessary, or that it explains the measured gains.

Statistical scope matters independently of accuracy. New-root intervals condition on three fitted pairs, and crossed-seed sensitivity remains inconclusive. Public-600 fallback is observed-panel development. Hard-110 adds frozen-fit, harder same-domain transfer, not cross-domain validation. Its later native-feedback comparison uses the already observed tasks. Exact-identity checks do not certify pretraining cleanliness. VALUE has the highest mean success among the systems evaluated on Hard-110; coverage of strong external methods remains insufficient for unrestricted SOTA.

The three formative cases do not calibrate false-stop rates, minimality or sufficiency. VMGR does not test AR versus iterative architecture. Negative variants, rank-16 results and nonmatched LEX 163/216 remain in the record.

8 Conclusion

Comparison diagnostics and method design serve the same question: which change does a gain identify? The six candidate gates separate blocked claims from reportable package outcomes in three formative cases. VMGR then fixes candidate support to test completion-cost supervision. VALUE gains 3.646 points on new exact identities and 10.606 points over same-search UNIFORM on Hard-110, with less measured inference work in the latter comparison. Search recovery explains a different part of the system's performance. The evidence supports that separation and a bounded planning improvement, not universal gate validity or an unrestricted SOTA claim.

AI USE STATEMENT

The human authors led the study and manuscript development, directing the research questions, algorithmic and experimental decisions, and interpretation. They drafted and revised the paper with AI assistance. Under their direction, OpenAI Codex assisted with proposing and refining ideas, implementing and debugging methods, constructing and checking synthetic-task data, orchestrating experiments, analyzing recorded results, retrieving literature, preparing figures and tables, and drafting and polishing sections. It also assisted with formulating and checking mathematical claims and derivations in Section 3 and Appendix A.1. External AI-generated review comments informed revision. Numerical results come from recorded experiments, not language-model estimates. The human authors take responsibility for all submitted text, claims, references, code and artifacts, including AI-assisted content.

REPRODUCIBILITY STATEMENT

The main text specifies the method, fixed training recipe, task populations and conditional analyses. Appendices A–F give implementation details, subgroup counts, resources, failures, replay boundaries and formative diagnostics. Two local CPU review archives are prepared: one for method checks, and one for paired binary outcomes and the original statistical functions. Appendix A specifies their contents, commands and limits. They do not reproduce full model training or generation. Anonymous supplementary delivery remains subject to author approval; no public release or existing reviewer access is claimed.

ETHICS STATEMENT

The studies use executable code and synthetic symbolic tasks rather than human-subject data. Success in small planning domains does not establish safe deployment in consequential real-world settings.

REFERENCES

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://jmlr.org/papers/v23/20-1335.html.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.507. URL https://aclanthology.org/2023.emnlp-main.507/.

HKUNLP. diffusion-vs-ar: Official implementation for beyond autoregression. GitHub repository, 2025. URL https://github.com/HKUNLP/diffusion-vs-ar/blob/6743981a4ba42062c95279e590f3991de3985581/scripts/path/train-mdm.sh. Audited revision 6743981a4ba42062c95279e590f3991de3985581, scripts/path/train-mdm.sh.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. Position: LLMs can't plan, but can help planning in LLM-Modulo frameworks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22895–22907. PMLR, 2024. URL https://proceedings.mlr.press/v235/kambhampati24a.html.

Piotr Kozakowski, Łukasz Kaiser, Henryk Michalewski, Afroz Mohiuddin, and Katarzyna Kańska. Q-value weighted regression: Reinforcement learning with limited data. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2022. doi: 10.1109/IJCNN55064.2022.9892633. URL https://arxiv.org/abs/2102.06782.

Zhihong Liu, Yang Li, Rengming Huang, Cewu Lu, and Panpan Cai. Any house any task: Scalable long-horizon planning for abstract human tasks. arXiv:2602.12244, 2026. URL https://arxiv.org/html/2602.12244v1. Version 1.

Parsa Mazaheri. REPOT: Recoverable program-of-thought via checkpoint repair. arXiv:2605.30052, 2026. URL https://arxiv.org/html/2605.30052v1.

Mohammad Norouzi, Samy Bengio, Zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. Reward augmented maximum likelihood for neural structured prediction. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/2f885d0fbe2e131bfc9d98363e55d1d4-Abstract.html.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. arXiv:1910.00177, 2019. URL https://arxiv.org/abs/1910.00177.

Qwen et al. Qwen2.5 technical report. arXiv:2412.15115, 2024. URL https://arxiv.org/abs/2412.15115v1. Version 1.

Stéphane Ross and J. Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. arXiv:1406.5979, 2014. URL https://arxiv.org/abs/1406.5979.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Xiangli Shi, Xiaomeng Zhu, Ye Tian, Yuchun Guo, Ziyang Sun, Lujie Yin, Yuxuan Zhou, and Yufei Huang. Plan right, then plan tight: Symbolic RL for efficient embodied reasoning. arXiv:2606.31260, 2026. URL https://arxiv.org/html/2606.31260v1.

Sam Toyer, Felipe Trevizan, Sylvie Thiébaux, and Lexing Xie. Action schema networks: Generalised policies with deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.12089. URL https://ojs.aaai.org/index.php/AAAI/article/view/12089.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Advances in Neural Information Processing Systems*, volume 36, pp. 38975–38987. Curran Associates, Inc., 2023. doi: 10.52202/075280-1693. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7a92bcdede88c7afd108072faf5485c8-Abstract.html. Datasets and Benchmarks Track.

Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. AlphaZero-like tree-search can guide large language model decoding and training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49890–49920. PMLR, 2024. URL https://proceedings.mlr.press/v235/wan24c.html.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822. Curran Associates, Inc., 2023. doi: 10.52202/075280-0517. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Beyond autoregression: Discrete diffusion for complex reasoning and planning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NRYgUzSPZz. Implementation details: arXiv:2410.14157v2, Appendix C.2.

Guojian Zhan, Likun Wang, Xiangteng Zhang, Jiaxin Gao, Masayoshi Tomizuka, and Shengbo Eben Li. Bootstrap off-policy with world model. In *Advances in Neural Information Processing Systems*, volume 38, pp. 134093–134121. Curran Associates, Inc., 2025. doi: 10.52202/085713-4470. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c290dcfbb071656287ee373cdbe8a259-Abstract-Conference.html.

Guanhua Zhang and Moritz Hardt. Inherent trade-offs between diversity and stability in multi-task benchmarks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58984–59002. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24u.html.

A IMPLEMENTATION AND REPRODUCIBILITY

The frozen optimizer is AdamW with betas (0.9, 0.999), epsilon 10^{-8} and weight decay 0.01. Gradient-norm clipping is 1.0. Sequences have at most 256 tokens; physical microbatches contain at most four candidates. LoRA targets the query, key, value and output projections, plus the gate, up and down MLP projections. Training uses eager attention and non-reentrant gradient checkpointing.

GradScaler starts at 4096. Its growth factor is two, backoff factor 0.5 and growth interval 2,147,483,647. The pinned runtime is PyTorch 2.4.1+cu121, Transformers 4.38.2 and PEFT 0.10.0. Table E1 lists all initialization seeds. The names FIT and TRAIN-selection distinguish fitting and development data.

Across its three fits, each policy processes 622,248 candidate sequences, 155,749 forward batches, 101,415,996 active input tokens and 9,333,720 supervised tokens. Checkpoints at updates 40, 80, 160 and 176 are retained. They are not reselected on the new-root panel.

Two local CPU review archives are prepared.

Method checks: extract `vmgr_cpu_method_review_v3.zip`, then run:

```
PYTHONPATH=src CUDA_VISIBLE_DEVICES= python -m pytest -q tests
```

This checks graph construction, targets and loss implementation on fixtures. CPU PyTorch is needed for the optional tensor tests.

Statistical reproduction: extract `vmgr_cpu_results_review_v1.zip`. With Python 3.9+ and NumPy, run:

```
python analyze.py
python -m unittest -v test_analysis
```

The analysis recomputes frozen statistics from all 4,608 new-root and 990 original Hard-110 binary outcomes, retaining task/fit pairing. Ten tests check exact fields and floating agreement within 10^{-10}, including adverse sensitivity findings. The result package needs no model, GPU, network or project checkout. Original task IDs, text, states and plans are omitted; public ordering could still permit linkage. These checks do not independently validate source outcomes or reproduce training, generation and cost replay. Anonymous delivery and release permissions remain subject to author approval; neither existing reviewer access nor full-model reproduction is claimed.

A.1 TARGET IDENTITIES AND LIMITATIONS

Write p_τ for normalized $\exp(-Q/\tau)$ and u for the uniform distribution. Equation (2) is $w = \epsilon u + (1-\epsilon)p_\tau$. Any candidate-independent baseline $B(x)$ cancels between numerator and denominator. Thus an advantage-style target using $Q - B(x)$ is identical unless clipping, rescaling or another transformation is also introduced. The implementation subtracts the minimum for stability.

If m candidates tie at minimum cost, the limit as τ approaches zero assigns $\frac{\epsilon}{K} + \frac{1-\epsilon}{m}$ to each minimizer and $\frac{\epsilon}{K}$ to every other candidate. As τ tends to infinity, or at $\epsilon = 1$, the target becomes uniform. At finite positive τ, the zero-floor limit remains strictly positive in exact arithmetic; the current executable contract excludes exact $\epsilon = 0$. Numerical underflow is a separate consideration. For every candidate, $\frac{\epsilon}{K} \le w \le 1 - \epsilon + \frac{\epsilon}{K}$. Moreover, $\mathrm{TV}(w, u) = (1-\epsilon)\,\mathrm{TV}(p_\tau, u)$, where TV is total variation distance. The floor therefore contracts deviation from uniform and provides a cost-independent lower bound. None of these bounds constrains learned policy probabilities.

The dynamic program evaluates complete supported paths. Its limitation is coverage. Suppose two candidates have observed completion costs eight and six. An absent continuation could let the first finish in four actions. The graph would prefer the second candidate even though the unrestricted optimum prefers the first. Exact transition verification cannot exclude this failure.

For a finite budget DAG, value evaluation is linear in nodes plus edges once transitions and topological order are available. Construction, ordering and execution replay have additional costs. Replacing the constant two in Equation (1) with a positive macro duration gives the corresponding

budget-decreasing recurrence. This mathematical extension is not evidence from longer-macro experiments.

CPU tests check baseline-shift invariance, floor bounds, total-variation contraction and limiting targets using synthetic cost vectors. They are implementation tests, not additional task evaluations or learned-model ablations.

A.2 TRAINING-INPUT LINEAGE AND GRAPH CONSTRUCTION

FIT denotes the 72 roots used for optimization. TRAIN-selection denotes a different set of 72 roots used repeatedly for recipe development. The frozen source construction uses problem seeds 20261629 and 20261630, respectively, with 12 roots per block count from six through eleven. Selection rows cannot enter the shared-support training builder.

For each fitting root, the earlier recurrent construction starts from the initial state and traverses all certified PLACE-subgoal continuations. It retains transitions that replay legally, end hand-empty, satisfy the selected placement goal, strictly reduce the number of unsatisfied goals, and stay within 36 primitive actions. Node identity includes state and committed action count. Every reached supported unsolved node is an anchor, not only the nodes along one chosen path. An anchor reached after c actions starts at budget $36 - c$. Unsupported unsolved nodes invalidate the source construction.

A constructive teacher supplies a canonical complete goal witness. A bounded search over legal first macros supplies alternative complete witnesses, with at most 400,000 expansions per first macro and at most 36 primitive actions. Expansion-censored or budget-infeasible branches cannot become supervision. The canonical witness is recovered from its certified constructive plan rather than accepted only if a second bounded search rediscovers it.

Each source constructor keeps the canonical witness plus up to three alternatives outside the canonical first-macro set. The lexical constructor orders alternatives by rendered first macro. The value-aligned constructor orders them by total plan length, residual goal count after the first macro, then rendered first macro. Missing alternative slots repeat the canonical witness. Every chosen plan is replayed to the exact goal and split into two-action transitions. Repetition in the old four-slot schedule is not a completion cost or an extra vote in the present objective.

VMGR uses only the two trajectory-closed source arms. It deduplicates exact targets at each root and serialized state/goal/budget prompt. It checks that root identity preserves the block universe and goal. A terminal macro has cost two. Every nonterminal macro must point to an observed successor with two fewer actions remaining. This prevents an unverified continuation from becoming a finite cost. Construction aborts on an invalid edge rather than dropping its root.

The union has 6,657 nodes and 9,428 edges. Of these nodes, 5,750 are singletons, 73 have multiple candidates with tied costs, and 834 have unequal costs. The last group is 12.528% of nodes but 11.643% of root-balanced objective mass. Those denominators should not be interchanged.

Algorithm A1: shared-support target construction.

1. Parse both completed witness corpora. Keep their distinct $(r, \mathrm{prompt}, \mathrm{target})$ triples.
2. Replay each target. Map every nonterminal edge to the exact successor prompt at budget $b - 2$; require that successor in the union.
3. Visit nodes in increasing budget. Compute Equation (1) from the already evaluated successors and terminal goal states.
4. Compute uniform or floor-softmin weights on the same sorted candidate list. Assign node coefficient $1/(|R||X_r|)$.
5. Accumulate every weighted suffix loss before each nine-root update. Use the same fitting schedule for all weighting policies.

All graph operations use fitting data. Deployment generates macro text from its prompt; it never accesses this graph or a stored completion witness. The finite-support guarantee concerns the input graph, not all states that the learned policy may visit.

A.3 POPULATION, ANALYSIS AND REPLAY DETAILS

The final new-root generator uses seed 20262151 for 6,144 candidates over blocks 6–11. In generator order, retain the first 128 per block that are initially unsolved and exact-identity-disjoint from earlier accessible history and already retained tasks. The exclusions also contain the entire earlier 6,144-candidate pool generated with seed 20262051. Insufficient supply stops construction without a new seed or top-up. These are the original frozen settings, not a newly sampled population.

The 768-root analysis uses PCG64DXSM seed 20262052, 20,000 draws and linear percentile quantiles. Resampling occurs within each of six block strata. Each root carries all three paired fit outcomes through the bootstrap. The estimand averages root differences within blocks, then weights blocks equally. Its block differences, for six through eleven blocks, are 3.385, 3.906, 1.302, 3.385, 5.729 and 4.167 percentage points. VALUE-only/UNIFORM-only counts by fit are 50/50, 73/46 and 97/40. These decompositions are descriptive, not six additional confirmatory tests.

For exact crossed-sensitivity reproduction, first draw a 20,000-by-3 matrix of fit indices, uniformly with replacement from indices 0–2, using int32. Then visit blocks 6–11 in order. For each block draw a 20,000-by-128 matrix of root indices, uniformly with replacement from 0–127. The conditional replicate averages each sampled root over all three fits. The crossed replicate uses the sampled fit columns, reusing the same three fit indices across every root and block in that replicate. Both average within blocks, then across the six blocks. Linear 2.5/97.5 percentiles give the reported intervals. The reproduction package reruns this frozen analysis without changing the outcome grid or sampling plan.

The public task source is the LLMs-Planning repository at revision `fc638a1aff7df3fe7a1a1d289fa2c04cc24dc284`. It retains source order and all 600 selected-source tasks. No task is replaced based on difficulty or model outcomes. Public bootstrap analyses retain paired fit outcomes within each task and weight tasks equally, not blocks equally. The original VALUE–UNIFORM and Best-of-16 comparisons use seed 20260924. Native Modulo and best-first use 20260925; fallback uses 20260926. Hard-110 uses 20260927, with all three paired fits retained in each of its 110 task clusters and no block stratification. The later permutation probe returns to the stratified 768-root analysis with seed 20260923. Each uses 20,000 PCG64DXSM draws and linear quantiles. Intervals are not jointly adjusted for multiple comparisons.

Protocols were recorded locally before their respective evaluations. This is not independently timestamped public preregistration. New-root model outcomes were inspected only after all six cells and three queues completed, 4,608 records existed and workers had exited. Subsequent public controls were registered after prior public results were known. This chronology is retained when interpreting later comparisons.

Composition checks retained selected execution trajectories and recorded work. Historical discarded neural beam tokens are unavailable and were not regenerated. New best-first and fallback records include expanded and abandoned branches, allowing fuller replay of those runs. Engineering missingness aborts composition rather than count as a task failure. Calibration, sealed-test and consumed S2A raw panels were not opened for this revision.

A.4 NEW-ROOT FAILURE COUNTS AND MEASURED WORK

UNIFORM has 346 action-budget and 85 no-valid-noncycle terminal failures. VALUE has 286 and 61. Candidate-level replay failures increase from 43,215 to 50,623, while parse failures decrease from 46,292 to 40,161. Candidate counts are not failed-task counts or identified causal mediators.

Resource across three fits	UNIFORM	VALUE
Reused training updates	528	528
Reused padded training token positions	159,295,488	159,295,488
Reused training cell-seconds	73,935.524	75,672.049
New generation calls	21,800	20,961
New prefill + decode token-work	57,443,056	55,243,536
New macro replay calls	294,609	290,931
New inference cell-seconds	33,190.499	33,237.918

Table A1. The original complete new-root resource accounting. Training is reused, not free. No fitting occurs during this evaluation.

The three concurrent queues consume 66,456.038 summed seconds including loading. The longest lasts 23,247.691 seconds, approximately 6 hours 27 minutes. Inference token-work falls by 3.829%, but measured cell wall time does not. These are observed implementation costs, not matched total FLOPs or a controlled hardware speedup. Construction and earlier research are additional.

A.5 Does the graph add information beyond recorded suffix lengths?

This diagnostic compares two cost constructions, Q_G and Q_W, on the same FIT graph. It is not the VALUE–UNIFORM experiment. VALUE uses cost-dependent targets, whereas UNIFORM ignores costs; their target weights differ at all 834 unequal-cost nodes. The 357-node and one-argmin counts below instead quantify what changes when graph costs replace witness costs inside the same floor-softmin formula.

A post-hoc structural diagnostic recovers complete witnesses from the original 3,923 FIT rows, using their anchor/slot metadata and exact cyclic-lane recovery. It recovers 2,857 distinct within-anchor witnesses across 509 anchors and 72 roots. Every recovered macro is replayed. Cyclic padding is checked but is not counted as additional path length. For each exact root/prompt/target edge, Q_W is the minimum remaining primitive-action length among all recovered witness occurrences. It uses the same budget-indexed identity as Q_G. The reconstructed graph, candidate ordering, costs and target weights exactly match the original prepared training input. No new planner or model is run.

FIT structural quantity	Observed value
Graph nodes / edges	6,657 / 9,428
Edges with $Q_G < Q_W$	390 (4.137%)
Nodes / roots containing such an edge	390 / 42
Root-balanced mass with a cost change	4.338%
Mean / maximum cost saving on changed edges, actions	4.754 / 14
Nodes with changed pairwise order or ties	174
Nodes with a changed minimum-cost candidate set	1
Nodes with changed floor-softmin weights	357
Root-balanced mass with changed weights	4.065%
Root-balanced target total variation	0.00194943

Table A2. Shared-graph costs versus minimum recorded witness-suffix costs. Mass weights roots equally and distinct prompts equally within each root. A target change requires total variation above 10^{-12}. This is a post-hoc FIT census, not a model-effect experiment or an evaluation set.

All 9,428 costs satisfy $Q_G \leq Q_W$, as required because each recorded suffix is a graph path. The strict improvements establish that the graph does more than relabel witness lengths. Their concentration and the single argmin change also constrain the contribution: most of its extra information refines relative costs and soft weights, rather than replacing the cheapest action. The diagnostic does not show that these 357 target changes caused the task gains. A model trained with Q_W would be a separate, unperformed ablation. The diagnostic uses no evaluation outcomes, no new optimizer updates and no model calls. Existing training data and all performance summaries remain unchanged.

B DEVELOPMENT-BASELINE COSTS

B.1 RETAINED DEVELOPMENTAL RESULTS

The repeatedly used 72-root panel is distinct from the final 768-root comparison. The selected rank-64 VALUE model solves 185/216 executions versus 180 for matched UNIFORM. Its 2.315-point gain has a descriptive root-cluster interval of $[-2.315, 6.944]$. Per-fit differences are +1.389, +6.944 and -1.389 points. Development is inconclusive on this contrast.

Development-only operating point	Successes /216	Success, %
Stepwise SFT, greedy	36	16.667
Complete-plan SFT, greedy	47	21.759
Complete-plan SFT, Best-of-16	114	52.778
Shared-support UNIFORM, beam 16	180	83.333
Shared-support VALUE, beam 16	185	85.648

Table B1. All methods use the same reused 72 roots and three fixed fits. They do not share training exposure or actual inference compute.

VALUE exceeds Best-of-16 by 32.870 points, with 75 paired wins and four losses. UNIFORM already exceeds it by 30.556 points. The full system gap therefore cannot be attributed to value weighting. Adapter ranks 16, 32, 64 and 128 scored 175, 182, 185 and 181 out of 216 in the final capacity-development comparison. This scan selected rank 64; it did not establish global optimality, monotonic scaling or exhaustion of the model's potential.

B.2 BASELINE TRAINING AND INFERENCE COSTS

The SFT label "4M" denotes a fixed target of four million active input tokens per fit, not four million examples or supervised output tokens. Full updates can overshoot the target. The three stepwise fits completed 3,768 updates and 12,037,173 active input tokens; the three complete-plan fits completed 4,560 updates and 12,019,899 active input tokens. Their summed cell wall times were 15,916.734 and 18,694.931 seconds. Eight and one additional AMP overflow attempts, respectively, are included in the recorded physical work. Resource qualification, precision-repair studies, and the abandoned failed-v1 partial work remain additional costs; the latter is only a recorded lower bound, not an exact historical total.

Best-of 16 reused the complete-plan fits without new optimization. Its 216 development episodes required 3,456 model calls, 814,512 prefill token-work positions, 247,113 decode token-work positions, and 23,581.697 cell-seconds. All generated candidates, including rejected ones, remain charged. The verifier recorded 2,066 calls and 1,329 cache hits, with 18,480 attempted and 16,837 applied simulator actions. The cache is part of this baseline's declared implementation; it is not silently treated as identical to VMGR's verifier.

The corresponding fixed VMGR development runs used 5,050,368 prefill-plus-decode positions and 2,851.698 cell-seconds. Thus VMGR used approximately 4.76 times the measured inference token-work but less elapsed time than this sequential Best-of 16 implementation. This is a hardware/implementation-dependent tradeoff, not equal-compute superiority. Candidate count and beam width are different resource units. Neither development comparison establishes a complete success–compute frontier on fresh problems.

C PUBLIC SEED AND DIFFICULTY BREAKDOWNS

The public complete-plan Best-of-16 control uses separate 4M-token SFT fits. It samples 16 plans at temperature 0.7, top-$p = 0.95$, top-$k = 50$, with a 256-token output cap and 1,024-token prompt cap. It verifies every candidate and selects the first success, otherwise the first legal plan. All sampled candidates are charged even if an earlier one succeeds. There is no outcome-driven checkpoint selection or baseline hyperparameter search.

The principal closed contrasts against the original VALUE decoder are:

Comparator	VALUE advantage, pp [95% interval]
UNIFORM	+2.556 [1.389, 3.778]
SFT best-first	−1.389 [−3.000, 0.167]
SFT Best-of-16	+13.667 [11.278, 16.056]
Fitted native Modulo	+39.556 [36.722, 42.444]
Verified-step SFT	+42.611 [39.444, 45.778]

Table C1. Conditional task-cluster intervals on 600 public tasks and three fixed fits. Different operating points are not compute-matched comparisons. The original negative best-first contrast is preserved.

VALUE has 78 paired wins and 32 losses against UNIFORM. Its per-fit increments are +2.500, +4.667 and +0.500 points. Against Best-of-16, the paired counts are 308 wins and 62 losses; fit differences are +12.667, +6.167 and +22.167 points. The large system difference includes different targets, training exposure and verifier interaction.

Seed	VALUE	UNIFORM	Complete-plan BoN16	Verified-step SFT	Best-first SFT
20261633	587	572	511	350	570
20261634	564	536	527	270	592
20261635	564	561	431	328	578

Table C2. Successes out of 600 for each fixed fit. All methods evaluate the same tasks; SFT columns represent different trained target formats.

Blocks	Seed	Tasks	VALUE	UNIFORM	BoN16	Verified-step SFT	Best-first SFT
3	20261633	100	98	95	88	61	100
3	20261634	100	97	90	96	49	100
3	20261635	100	98	94	72	65	100
4	20261633	445	436	425	375	257	434
4	20261634	445	418	397	383	198	444
4	20261635	445	413	416	313	234	438
5	20261633	55	53	52	48	32	36
5	20261634	55	49	49	48	23	48
5	20261635	55	53	51	46	29	40

Table C3. Unpooled public counts. VALUE−UNIFORM is negative for four blocks, seed 20261635 (−0.674 points), and zero for five blocks, seed 20261634. Positive aggregate differences do not imply improvement in every cell.

Leaving out seeds 20261633/34/35, respectively, gives VALUE−UNIFORM differences of +2.583/+1.500/+3.583 points and VALUE−BoN16 differences of +14.167/+17.417/+9.417 points. These are fixed-fit sensitivity summaries, not additional independent experiments. Public root exchangeability is assumed by the bootstrap, not established by observing positive intervals.

For BoN16, prefixes of 1, 4, 8 and 16 candidates yield 876, 1,261, 1,397 and 1,469 successes, respectively, out of 1,800 episodes. These prefixes are descriptive; N= 16 was fixed rather than chosen by these outcomes. No sampled candidate reaches the 256-token cap, but this alone does not establish an optimal prompt or the baseline's maximum attainable score.

D Public resources, failures and replay scope

The complete-plan fits include one AMP overflow: 4,561 physical attempts, 41,049 forward batches, 12,022,591 physical input tokens and 2,259,549 physical supervised tokens. Table D1's token counts describe successful updates; these inclusive physical totals prevent hiding the failed attempt; they

Recorded work across three fits	VALUE	UNIFORM	Complete-plan SFT + BoN16
Reused training updates	528	528	4,560
Reused active training input tokens	101,415,996	101,415,996	12,019,899
Reused supervised training tokens	9,333,720	9,333,720	2,259,072
Reused training worker-hours	21.020	20.538	5.193
Recorded public generation calls	6,737	6,796	28,800
Prefill token-work	11,660,960	11,758,032	5,672,544
Decode token-work	1,568,752	1,598,256	1,316,962
Returned tokens	1,613,500	1,640,006	1,345,762
Recorded public inference worker-hours	2.964	3.135	36.088

Table D1. Final-fit and public-inference costs, not all-in research costs. Token-work counts and summed worker wall time are distinct measures; neither is FLOPs. Reference training is charged once even when its checkpoints are reused in several evaluation panels. Retained VALUE public inference is also counted once across these contrasts.

are not added again to successful-work totals. The stepwise SFT training costs and its eight overflow attempts are reported in Appendix B. Neither training updates nor active-token exposure is matched across SFT and VMGR. Comparing 528 with 4,560 updates without their token counts is misleading.

Public verified-step SFT used 13,225 generations, 64,376,592 prefill and 2,104,928 decode positions, and 17,078.736 summed cell-worker seconds (4.744 hours). That total includes aborted-cell time with retained useful episodes; it is not all discarded failure overhead.

BoN16 takes 129,915.501 summed inference seconds versus 10,670.744 for VALUE. However, VALUE uses 1.893 times BoN16's counted prefill-plus-decode positions. The apparent 12.175-fold worker-time difference is implementation- and hardware-dependent; it is not a causal speedup at equal compute. Sequential sampling, beam execution, caching and differing plan representations prevent equating their cost units. Recorded model, generation and verifier sub-times are contained within total worker time and must not be added again.

BoN16 charges all 28,800 candidates: 14,187 achieve the goal, 13,381 contain illegal actions, 985 are legal without reaching the goal, and 247 fail parsing. Of 331 unsolved episodes, 270 have no legal plan and 61 have a legal but unsuccessful plan. VALUE has 85 unsolved episodes: 83 no-valid-noncycle terminations and 2 action-budget terminations. Failure taxonomies differ across decoders and do not by themselves identify a causal mechanism. BoN verification uses 8,835 calls and 19,718 cache hits, with 58,937 attempted and 53,217 applied simulator actions; cached and rejected work remain recorded.

BoN qualification adds 20 model calls and 74.472 seconds; its CPU composition takes 24.301 seconds with no new model calls. UNIFORM qualification adds 10 calls and 62.068 seconds. VALUE training qualification adds 3 updates and 362.487 seconds. Historical graph construction, precision qualification, failed launch and recovery work, canceled training and development searches remain additional. Known costs are retained in the artifact ledger; unavailable historical totals are unknown, not zero. Recovered SFT inference combines 1,179 immutable v2 receipts with 621 v3 receipts, rather than discarding failed engineering attempts or repeating all 1,800 episodes. These engineering versions are not independent scientific replications.

Replay checks all retained BoN candidate text/tokens, verifier outcomes and recorded work. For VALUE and UNIFORM it checks retained selected trajectories and available work accounting. Discarded neural beam-candidate tokens are not available for complete replay; the paper does not claim they were regenerated or exhaustively checked. No public reference plans are used for selection.

D.1 FITTED NATIVE-FEEDBACK CONTROL

This adaptation reuses the three complete-plan SFT fits used by Best-of 16, without retraining or checkpoint selection. Greedy batch-one decoding permits 15 rounds, 1,024 new tokens per call and 32,768 context tokens. Each round starts its plan from the original initial state. The licensed author iteration loop receives native-symbol first-error VAL feedback and retains the full raw history. It stops at a verified success, null output or context exhaustion. The qualified bounded-eager backend changes memory execution, not this scientific contract. It is an adaptation, not a reproduction of the author's model recipe.

Native Modulo improves Direct by 2.500 points, with conditional task-cluster 95% interval [1.778, 3.278]. Direct is the included first-attempt subset. Success retention makes its zero paired losses structural. The extra rounds also add compute, so the contrast does not isolate the information in feedback.

Seed	V /600	M /600	D /600	V−M, pp (wins/losses)	M−D, pp (wins/losses)
20261633	587	360	352	+37.833 (230/3)	+1.333 (8/0)
20261634	564	378	369	+31.000 (205/19)	+1.500 (9/0)
20261635	564	265	237	+49.833 (305/6)	+4.667 (28/0)
Total /1,800	1,715	1,003	958	+39.556 (740/28)	+2.500 (45/0)

Table D2. All fixed fits. V denotes VALUE, M native Modulo, and D Direct. Direct is each native episode's first attempt, not a candidate from the earlier sampled Best-of 16 experiment.

Blocks	Seed	V	M	D	V−M, pp	M−D, pp
3 (100 tasks)	20261633	98	75	72	+23.000	+3.000
3	20261634	97	69	67	+28.000	+2.000
3	20261635	98	54	52	+44.000	+2.000
4 (445 tasks)	20261633	436	256	251	+40.449	+1.124
4	20261634	418	279	275	+31.236	+0.899
4	20261635	413	186	162	+51.011	+5.393
5 (55 tasks)	20261633	53	29	29	+43.636	+0.000
5	20261634	49	30	27	+34.545	+5.455
5	20261635	53	25	23	+50.909	+3.636

Table D3. All block-by-fit cells. V, M and D follow Table D2; entries are success counts, not filtered subsets.

Aggregating the three fits within blocks 3/4/5, VALUE−Modulo is +31.667, +40.899 and +43.030 points, with paired wins/losses 97/2, 569/23 and 74/3. Modulo−Direct is +2.333, +2.472 and +3.030 points, with 7/0, 33/0 and 5/0. Removing fits 20261633, 20261634 and 20261635 in turn yields primary estimates +40.417, +43.833, +34.417; secondary estimates are +3.083, +3.000, +1.417 points. These are conditional sensitivity summaries, not additional independent trials. The task-cluster interval assumes task exchangeability and conditions on six fixed fits; it does not estimate uncertainty across the population of possible training runs.

Native Modulo performs 13,085 generations, 10,715,022 prefill positions and 350,543 decode positions, returning 363,628 tokens. Summed worker time including loading is 36,327.137 seconds (10.091 hours), not the parallel elapsed time. Model time 36,111.278 seconds and verifier time 90.963 seconds are contained components, not additional charges. Direct's 1,800 calls, 354,534 prefill positions, 81,063 decode positions and 82,863 generated tokens are an included subset. No additional Direct or VALUE inference is performed. The complete-plan training in Table D1 is reused, not free; neither it nor retained VALUE training/inference is counted again as new physical work. No new training or model calls occur in composition.

Of the 13,085 native rounds, 11,067 contain illegal actions, 645 fail format checks, 370 are legal without reaching the goal, and 1,003 succeed. All 797 unsolved episodes reach the 15-round budget.

This describes a full-plan generation/revision bottleneck; it does not causally identify a training or architecture repair. Native verification uses 12,440 VAL calls, with 25,637 attempted and 14,570 applied actions in the separately recorded exact simulator. VAL's internal primitive-transition cost is unavailable, not equal to that simulator count and not zero. VALUE's 85 episode failures remain 83 no-valid-noncycle-candidate and 2 action-budget failures. The two failure taxonomies are not interchangeable.

Backend and interface history adds 47 generation calls, 10 completed stress forwards and one failed stress attempt, including the retained context-lifecycle, serialization-comparator, SDPA numerical and mask-allocation failures. The final bounded backend qualification contributes 9 of those calls and 6 of those stress forwards; its 114.984 summed worker seconds are included in that history, not an extra duplicate qualification. Fixed-fixture token/logit checks and full 32K feasibility do not imply universal bitwise eager equivalence. Earlier research, CPU preflight and missing historical cost totals remain additional and unknown.

CPU composition takes 124.559 seconds separately from original inference. It replays all 1,800 new token histories, author-loop decisions and actual VAL outputs, plus 1,800 retained VALUE selected trajectories and aggregate work. It regenerates no model outputs. The observed public panel, different training exposure, whole-plan versus two-action representation, feedback and inference budgets remain limitations. The scoped comparison is neither compute-matched nor independent confirmation or unrestricted public SOTA.

D.2 Best-first search: complete replay and resource tradeoff

The best-first frontier reuses the fixed stepwise SFT fits. It generates primitive actions with beam 16 and retains per-branch histories. Path-local cycles are rejected. The priority is the sum of native length-normalized beam scores, not path log probability. Every generated legal child receives terminal checking. The limits are 36 total expansions and 36 primitive actions per path, with 128 new tokens and a 1,024-token prompt cap per call. Every expanded and abandoned branch remains in the cost accounting.

All 1,800 best-first episodes completed, with no engineering failures. Full CPU composition replays expanded and abandoned branches, physical beam work, the retained 1,800 receding-horizon SFT trajectories (the 1,179 + 621 historical union), and 1,800 VALUE selected trajectories and aggregate work. It takes 64.537 seconds and generates no new model outputs. This replay cost is separate from inference.

Seed	V−BF, pp	Paired wins/losses	BF−SFT	Paired wins/losses
20261633	+2.833	28/11	+36.667	233/13
20261634	−4.667	7/35	+53.667	325/3
20261635	−2.333	20/34	+41.667	258/8
Pooled	−1.389	55/80	+44.000	816/24

Table D4. Fixed-fit paired contrasts. Primary conditional task-cluster 95% interval: [−3.000, 0.167] points; secondary interval: [40.722, 47.278].

Leaving out seeds 20261633/34/35 gives primary differences −3.500/+0.250/−0.917 points, and secondary differences +47.667/+39.167/+45.167. These are sensitivity analyses of three fixed fits, not new training-seed replications.

Blocks	VALUE / executions	Best-first / executions	V−BF, pp	W/L	BF−SFT	W/L
3	293/300	300/300	−2.333	0/7	+41.667	125/0
4	1267/1335	1316/1335	−3.670	17/66	+46.966	632/5
5	155/165	124/165	+18.788	38/7	+24.242	59/19

Table D5. V is VALUE; BF is best-first; W/L is paired wins/losses. Differences are in percentage points. All prespecified block strata, with unpooled seed counts in Table C3. The five-block contrast is descriptive; no subgroup-selected primary claim is made.

Best-first makes 23,704 generation calls and 260,720 model forwards. Its recorded work is 114,810,992 prefill positions, 3,792,256 decode positions and 3,956,713 returned tokens; no returned sequence reaches the generation cap. Summed worker time including loading is 30,616.662 seconds (8.505 worker-hours); the three workers take 10,984.785/9,287.144/10,344.733 seconds. Synchronized generation time 30,466.229 seconds and model-attempt wall time 30,499.599 seconds are included components, not additive costs. Against retained VALUE inference, best-first uses 8.965 times counted prefill-plus-decode work and 2.869 times summed worker time. These are implementation-specific descriptive ratios, not matched-compute speedups, FLOPs, or independent wall-clock experiments.

The reused stepwise fits cost 3,768 successful updates, 12,037,173 active input tokens and 356,166 supervised tokens. Including eight overflow attempts gives 3,776 physical attempts, 33,984 forwards, 12,062,462 input tokens and 356,935 supervised tokens. Training worker time is 15,916.734 seconds (4.421 hours), versus 21.020 hours for the retained VALUE fits, which consume 101,415,996 active input tokens and 9,333,720 supervised tokens. Fits and VALUE inference are reused, not free and not charged again for each paired contrast. Lower VMGR inference work therefore does not establish lower total training-plus-inference cost.

All 60 best-first unsuccessful episodes exhaust the expansion budget. Across 379,264 proposed candidates, the replay counts 109,729 parse failures, 190,909 illegal actions, 14,698 cycle rejections, 21,975 duplicates and 60 wrong-action-count outputs. There are 43,693 created nodes and goal checks, 247,500 simulator action attempts and 56,591 applied actions. VALUE retains 85 failures: 83 with no valid noncycle candidate and 2 exhausting the action budget. These failure categories describe different algorithms and are not interchangeable.

The new toy integration adds 4 calls, 42 forwards, 19,296 prefill positions, 608 decode positions and 660 returned tokens; its 12.457 seconds including loading contain 5.062 seconds of synchronized generation. Earlier public SFT decoder qualification retains 20 calls and one failed zero-call load, 54.085 seconds in total including 4.726 failed-load seconds. Earlier training, backend failures, and canceled-extension research remain additional historical costs in the retained records, including 37 completed canceled updates; incomplete-attempt and startup costs are unknown, not zero. Historical LEX 163/216 remains an unmatched reference rather than a baseline for these 600 public tasks.

The observed panel, unequal training and actual inference budgets, and primitive versus two-action representations limit this comparison. The secondary contrast also changes terminal checking and cannot isolate backtracking. Neither the negative aggregate point estimate nor its zero-overlapping interval establishes equivalence. No new effect threshold or outcome-driven search is introduced.

D.3 Primary-preserving macro fallback

The extension was frozen after observing the public comparisons. It preserves the original primary policy, including first-duplicate retention, tie order and path-local endpoint-cycle rejection. Legal siblings remain available only for fallback. Branch prompts use their own remaining action budget, not the number of calls spent on other branches. The frontier uses summed native length-normalized beam scores and stable insertion-order ties. There is no learned critic, goal-distance oracle, training-graph lookup or reference-plan access. The limit is 36 physical generation calls per episode and 36 primitive actions per path.

Every historical success is retained: 1,715 VALUE and 1,669 UNIFORM executions. All 85 VALUE and 131 UNIFORM failures are reconstructed without filtering. Their primary actions,

terminal states, failure reasons, calls and candidate / verifier counts match the originals. The final denominator is 600 public task clusters and three fixed fits per policy, not just the failed subset. Zero losses against each old policy follow from this reuse design; they do not constitute independent evidence that a deployment rerun can never regress.

Blocks	Seed	VALUE-fallback	UNIFORM-fallback	SFT best-first
3, n=100	20261633	100	100	100
3, n=100	20261634	100	100	100
3, n=100	20261635	100	100	100
4, n=445	20261633	445	445	434
4, n=445	20261634	445	440	444
4, n=445	20261635	445	445	438
5, n=55	20261633	55	55	36
5, n=55	20261634	54	54	48
5, n=55	20261635	55	55	40

Table D6. All block-by-fit successes. The two 54/55 macro outcomes in the second fit fail on different tasks: VALUE–UNIFORM has one win and one loss in that stratum.

VALUE-fallback minus SFT best-first is +5.000, +1.167 and +3.667 points, with paired wins/losses 30/0, 8/1 and 22/0. Blockwise differences are 0.000, +1.423 and +24.242 points, with paired counts 0/0, 19/0 and 41/1. Leave-one-fit-out means are +2.417, +4.333 and +3.083 points. The enhanced VALUE–UNIFORM seed differences are 0.000, +0.833 and 0.000; corresponding leave-one-fit-out means are +0.417, 0.000 and +0.417 points. Its [0.000, 0.556] interval establishes neither a training-seed population benefit nor equivalence.

Relative to the original policy, VALUE recovers 84 failures for +4.667 points [3.389, 6.000]; UNIFORM recovers 125 for +6.944 points [5.333, 8.611]. The one remaining VALUE failure and six UNIFORM failures exhaust the expansion budget. All four intervals use 20,000 PCG64DXSM task-cluster draws, seed 20260926, and are conditional on fixed fits and the observed-panel reuse protocol. There is no outcome-size gate or multiplicity-adjusted general superiority claim.

Recorded work	VALUE + fallback	UNIFORM + fallback	SFT best-first
Generation calls	7,561	8,137	23,704
Prefill + decode token-work	14,904,304	16,062,960	118,603,248
Inference worker-hours	3.311	3.697	8.505
Reused training worker-hours	21.020	20.538	4.421

Table D7. Actual original-plus-new inference accounting includes repeated primary prefixes. Original inference is charged once, not once per contrast. Qualification, CPU replay and other historical research remain additional. These are unequal-resource operating points, not matched-compute speedups.

The new work comprises 2,165 model calls: 915 reconstruct primary prefixes and 1,250 expand fallback branches. Their respective prefill-plus-decode counts are 1,841,696 and 2,539,568. VALUE contributes 824 new calls and UNIFORM 1,341. The three queues consume 3,271.194 summed worker seconds, with the longest taking 1,926.300 seconds. Summed generation time, 3,215.455 seconds, is an included component, not another charge. No optimizer updates are added. GPU qualification adds five physical calls and 13.498 seconds, with four explicitly synthetic routing calls excluded from model-call counts. There are zero new engineering failures; historical failures and unknown costs remain.

CPU composition takes 10.320 seconds and generates no model outputs. It replays every new expanded or abandoned branch, beam tokens, scores and physical work; it validates old selected trajectories and available aggregate work, not unavailable historical discarded beams. The already closed

SFT comparison vector reproduces its original summary. The 7.958-fold difference in counted SFT versus VALUE-fallback inference token-work does not establish lower all-in VMGR cost: its reused training is substantially larger. Excluding reconstructed prefixes would be a counterfactual deployment ledger, not observed wall-time savings.

D.4 FROZEN-FIT HARD-110 TRANSFER

Population and chronology. We retain all 110 tasks, in source order, from the hard Blocksworld configuration at the public revision given in Appendix A.3. The configuration, fits, decoding budgets and analysis were fixed after the smaller-panel fallback study and before inspecting Hard-110 outcomes. No reference plans were supplied to the models or read for endpoint selection. This is prospective evaluation on a harder same-domain panel, not a new fit, cross-domain test or independently timestamped public preregistration.

The block-count distribution is 1, 1, 7, 14, 13, 12, 13, 18, 21 and 10 for 6 through 15 blocks. Exact-identity comparison finds zero overlap with the previous 600-task public panel and zero with 2,682 unique fitting prompt states represented by 6,657 training rows. It does not certify non-isomorphism, pretraining cleanliness or exclusion from every historical development set. Forty-nine tasks require object names beyond the legacy input alphabet a–l. An explicit a–z input option admits them; existing action parsers already support these names. An input-only constructive upper bound is at most 48 actions. This is not a shortest-path distance or a plan supplied at inference.

Frozen operating points. VALUE and UNIFORM reuse their rank-64/alpha-128 update-176 fits. SFT reuses the three fixed four-million-input-token stepwise fits. There is no retraining, checkpoint selection, root replacement or HPO. All methods use ordinary beam 16, at most 128 generation calls and 64 primitive actions per path. Macro fallback generates at most 16 new tokens per call. SFT uses native-chat single-action prompts, a 128-new-token and 4,096-prompt-token cap, and at most 128 best-first expansions. Prompts are not truncated. The base model is FP16 and adapters are FP32. Exact transition execution and terminal goal satisfaction determine success. Shared nominal call limits do not make different representations or realized computation equivalent.

Complete results. The primary comparison is VALUE-fallback minus SFT best-first; same-search VALUE minus UNIFORM is secondary. Both use 20,000 unstratified PCG64DXSM task-cluster draws, seed 20260927, linear percentile quantiles and the three fixed pairs within each resampled task. The intervals are conditional task uncertainty under exchangeability, not a training-seed population test. There is no added effect threshold, all-positive-seed gate, optional stopping or multiplicity-adjusted general superiority claim.

Blocks	Tasks	Fit 20261633, V/U/S	Fit 20261634, V/U/S	Fit 20261635, V/U/S
6	1	1/1/0	1/1/1	1/1/1
7	1	1/1/0	1/1/0	1/1/1
8	7	7/6/0	7/6/1	7/6/4
9	14	14/14/0	13/12/1	13/13/1
10	13	13/11/1	8/7/1	10/9/1
11	12	7/9/0	7/7/0	11/8/0
12	13	6/8/0	7/3/0	4/8/0
13	18	8/7/0	11/6/0	11/7/0
14	21	11/6/0	5/5/0	12/5/0
15	10	4/1/0	1/1/0	4/1/0
All	110	72/64/1	61/49/4	74/59/8

Table D8. Every Hard-110 block-by-fit cell. Each triple gives success counts for VALUE, UNIFORM and SFT, with the task count as each denominator. All fixed fits and all negative subgroup differences are retained.

The primary difference is +58.788 points [52.121, 65.455], with 194 wins and zero losses. Per-fit paired wins/losses are 71/0, 57/0 and 66/0. Omitting fits 20261633, 20261634 and 20261635 gives +55.909, +62.273 and +58.182 points. The secondary difference is +10.606 points [4.848, 16.667], with 68 wins and 33 losses. Its per-fit paired counts are 18/10, 21/9 and 29/14; leave-one-fit-out means are +12.273, +10.455 and +9.091 points. VALUE–UNIFORM block means, from 6 to 15

blocks, are 0, 0, +14.286, +2.381, +10.256, +2.778, −5.128, +18.519, +19.048 and +20.000 points. These are descriptive subgroups, not separately confirmed discoveries. No crossed-seed population interval is claimed for this three-fit transfer study.

All 123 VALUE, 158 UNIFORM and 317 SFT unsuccessful executions terminate by expansion-budget exhaustion. SFT generates 666,576 candidate sequences, including 188,944 parse failures, 273,281 illegal-action rejections and 485 wrong-action-count rejections. VALUE and UNIFORM respectively record 22,999 and 28,733 parse failures, and 42,881 and 41,264 macro replay failures. These candidate diagnostics are not mutually interchangeable with failed tasks or isolated causes of the success gap. The SFT control's poor transfer limits what the system comparison establishes; it does not establish failure of all SFT, ToT or RAP implementations.

Physical work and recovery. The first attempt completed 505 unique episodes. A GPU availability guard stopped a handoff before loading its next model; the other workers stopped at task boundaries. Its six failure records remain immutable. There were no unfinished-root model calls. Recovery computes only the remaining 485 episodes, using 11 workers in six fixed GPU queues. The final replay reconciles all 990 unique task–fit–method records, every expanded or abandoned branch, model tokens, scores and recorded physical work. There are zero new recovery or composition failures. The partial attempt is retained work, not an extra 505 episodes added to the scientific denominator.

Hard-110 work across three fits	VALUE	UNIFORM	SFT best-first
Generation calls	20,879	24,236	41,661
Model forward calls	315,611	369,055	450,908
Prefill token-work	66,433,520	77,542,832	261,906,528
Decode token-work	4,715,712	5,517,104	6,547,952
Returned tokens	4,983,432	5,808,742	6,909,597
Length-capped candidate sequences	780	773	0
Model-generation seconds	33,010.965	38,466.871	56,055.593

Table D9. Complete original-plus-recovery inference work. The first attempt's completed episodes are charged once. Model-generation time is included in worker elapsed time, not added to it. Token-work is not FLOPs; generation timing is an implementation measurement, not a controlled hardware speedup.

Summed worker elapsed time, including loading, is 128,447.624 seconds (35.680 worker-hours). The retained first attempt accounts for 55,566.480 seconds and recovery for 72,881.144 seconds. There are 20 model loads: nine retained and 11 new recovery loads. These summed times are not parallel completion latency. New training updates are zero. Prior VALUE, UNIFORM and stepwise-SFT training remains 21.020, 20.538 and 4.421 worker-hours, respectively; the same fits are reused across public panels, not recharged as new training.

Backend qualification adds eight physical generation calls and 26.810 seconds; input preparation takes 9.393 seconds. CPU composition adds 465.445 seconds and zero model calls. These are separate preparation/replay costs, not another inference run. Exact failed-queue overhead is unknown and cannot be set to zero or added twice to included worker times. The historical qualification, failed attempts and research costs in Appendices B–E remain additional; their all-in total is incomplete. Faster counted inference does not erase those costs.

D.5 Native full-plan feedback on observed Hard tasks

The original Hard-110 results prompted a separate method-coverage extension. It retains all 110 tasks, three complete-plan SFT fits and the existing VALUE outcomes. Models are not retrained or selected again. The comparison was fixed before these new generations, but after the original Hard outcomes were known. It is an observed-panel follow-up, not another independent confirmation.

The control runs the retained author iteration kernel with our SFT prompt and symbol adapter, using actual VAL feedback. It follows the LLM-Modulo interface principle, not the author's complete model recipe (Kambhampati et al., 2024). Each fit uses its fixed four-million-token endpoint. Native greedy decoding uses batch one, 15 rounds, 1,024 new tokens per call, 32,768 context tokens and a

64-action plan ceiling. Raw history is retained. Each proposed plan restarts from the original state; malformed outputs consume a round without prefix salvage. The first attempt supplies Direct at no extra generation cost.

Native Hard fit	VALUE /110	Modulo /110	Direct /110	V−M, pp	V−M wins / losses
20261633	72	3	3	+62.727	69 / 0
20261634	61	0	0	+55.455	61 / 0
20261635	74	1	1	+66.364	73 / 0
Total /330	207	4	4	+61.515	203 / 0

Table D10. Complete fixed-fit native-feedback extension. Direct is the first attempt within the same episode, not an independently sampled baseline. Modulo adds no successful tasks in any fit. VALUE and native Modulo differ in training, action representation, feedback and inference resources.

The primary VALUE–Modulo difference is +61.515 points. Its conditional task-cluster 95% interval is [54.848, 68.182]. The analysis uses 20,000 unstratified PCG64DXSM bootstrap draws with seed 20260927, retaining all three fit outcomes in each of the 110 task clusters. Leave-one-fit-out means are +60.909, +64.545 and +59.091 points when omitting fits 20261633, 20261634 and 20261635, respectively. These are fixed-fit task comparisons, not inference over future training seeds. Task exchangeability remains an assumption.

The secondary Modulo–Direct difference is zero, with zero paired wins or losses. Its empirical bootstrap interval collapses to [0, 0] because every observed paired difference is zero. This records no observed rescue; it is not evidence that the population effect is exactly zero or an equivalence test. In general, the contrast includes extra computation as well as feedback.

Blocks	Tasks	Fit 20261633, V/M/D	Fit 20261634, V/M/D	Fit 20261635, V/M/D	V−M, pp
6	1	1 / 0 / 0	1 / 0 / 0	1 / 0 / 0	+100.000
7	1	1 / 0 / 0	1 / 0 / 0	1 / 0 / 0	+100.000
8	7	7 / 0 / 0	7 / 0 / 0	7 / 0 / 0	+100.000
9	14	14 / 2 / 2	13 / 0 / 0	13 / 1 / 1	+88.095
10	13	13 / 1 / 1	8 / 0 / 0	10 / 0 / 0	+76.923
11	12	7 / 0 / 0	7 / 0 / 0	11 / 0 / 0	+69.444
12	13	6 / 0 / 0	7 / 0 / 0	4 / 0 / 0	+43.590
13	18	8 / 0 / 0	11 / 0 / 0	11 / 0 / 0	+55.556
14	21	11 / 0 / 0	5 / 0 / 0	12 / 0 / 0	+44.444
15	10	4 / 0 / 0	1 / 0 / 0	4 / 0 / 0	+30.000

Table D11. Every Hard-native block stratum, without outcome-based filtering. V, M and D denote VALUE, native Modulo and its own Direct attempt. The M−D difference is zero in every cell. Each fit evaluates the stated task count; counts across three fits are not three independent task samples.

All 326 unsuccessful native episodes exhaust the 15-round limit. Across 4,894 rounds, 4,644 produce illegal actions, 202 fail formatting, 44 execute without reaching the goal, and four succeed. Format failures comprise 59 missing/invalid plan blocks, 115 unknown-object/operation/arity errors, 19 other arity errors, six malformed actions and three action-budget/noncanonical sequences. The reference VALUE failures are 123 expansion-budget exhaustions. Full history, token, author-loop and actual-VAL replay reconciles these counts. It establishes faithful execution of this adaptation, not strong transfer competence of the generator. This control's failure does not refute native feedback methods in general.

Native work across three fits	Terminal total	Direct subset
Model calls	4,894	330
Prefill token-work	5,664,773	87,666
Decode token-work	224,276	44,307
Generated tokens	229,170	44,637

Table D12. Newly measured native inference. Direct is included in the terminal total and must not be added again. The retained VALUE inference in Table D9 uses 71,149,232 prefill-plus-decode positions versus 5,889,049 here. Higher success therefore comes with more counted inference work against this control.

There are 4,692 VAL calls, 10,173 exact-simulator action attempts and 5,529 applied actions. VAL-internal primitive work is not measured by those simulator counts. Six disjoint 55-task shards require six model loads. Summed worker time is 23,191.404 seconds (6.442 worker-hours); it is not parallel latency. Model-call time is 23,054.169 seconds, observed generation time 23,041.719 seconds, and verifier time 36.421 seconds. These timing components overlap with worker time and are not additive.

New training updates are zero. Reused complete-plan training across the three fits comprises 4,560 updates, 4,561 attempts, 12,019,899 active input tokens and 12,022,591 physical input tokens. Its 18,694.931 seconds (5.193 worker-hours) remain part of the trained-system cost. VALUE's reused training remains 21.020 worker-hours. Neither is recharged as new work each time a panel is evaluated.

Twelve synthetic device checks add 4,068 prefill and 1,572 decode token-work positions, 1,584 generated tokens, six model loads and 216.851 worker seconds. They establish repeat and paired-device agreement on the checked inputs; success was not an admission criterion. Earlier three-fit long-context qualification is reused. Six initial wrapper imports failed before loading a model because the launcher omitted its module path. Their original logs remain; new model calls were zero, but unmeasured CPU startup time is unknown, not zero. After authorization, only the launch environment was corrected.

One CPU composition replays all 330 new episodes and the retained VALUE selected trajectories and aggregate work in 150.791 seconds, with zero model calls. The original Hard replay already checked its discarded branches; this follow-up does not independently replay them again. Prior 47 native-backend generation calls, ten completed stress forwards, one failed stress attempt and the research/failure costs in Appendices B–E remain additional context. The all-in research total is incomplete; reused or unknown costs are not zero. No new training, prompt search or budget sweep follows this outcome.

E MECHANISM SENSITIVITY AND HISTORICAL RESEARCH COST

Fixed fit	UNIFORM /768	VALUE /768	PERMUTED /768	V−U, pp	V−P, pp
20261633	656	656	650	0.000	+0.781
20261634	602	629	637	+3.516	−1.042
20261635	615	672	605	+7.422	+8.724
Total /2,304	1,873	1,957	1,892	+3.646	+2.821

Table E1. New-root VALUE–UNIFORM comparison and later PERMUTED control. V, U and P denote the three policies. There are 768 task clusters, not 2,304 independent tasks. The PERMUTED evaluation reuses the observed panel.

Measured work, three fits	VALUE + fallback	UNIFORM + fallback	SFT best-first
Reused training worker-hours	21.020	20.538	4.421
Public inference worker-hours	3.311	3.697	8.505
Public inference token-work, millions	14.904	16.063	118.603
Hard-110 generation calls	20,879	24,236	41,661
Hard-110 token-work, millions	71.149	83.060	268.454

Table E2. Success–resource tradeoff for the strongest public operating points. The first inference rows concern the smaller 600-task panel; Hard-110 is separate. Inference includes repeated primary prefixes. Worker-hours are summed elapsed times, not parallel latency or kernel-active time. Token-work is not FLOPs. Qualification, failed attempts and historical research are additional costs reported in Appendices B, D and E; unknown totals are not zero.

A later 7B external search comparison was cancelled for submission timing after four of 330 planned executions, with two additional interrupted ledgers. Its partial outcomes are excluded from all performance claims. Setup, qualification and interrupted inference still incurred costs outside these completed-study tables; their all-in total is not available and is not zero.

Table E3 restores the full sequence of matched weighting contrasts, including the earlier positive rank-16 mean and the inconclusive development comparison. It is an evidence map, not a meta-analysis or an exhaustive table of all historical model variants. Every row compares VALUE and UNIFORM within its own recipe; decoder and population differ between rows.

Weighting study	VALUE / UNIFORM successes	V−U, pp	Conditional 95% CI, pp
Rank 16, beam 4, earlier new roots	1,826 / 1,781 of 2,304	+1.953	[0.303, 3.602]
Rank 64, beam 16, reused development	185 / 180 of 216	+2.315	[−2.315, 6.944]
Rank 64, beam 16, final new roots	1,957 / 1,873 of 2,304	+3.646	[2.040, 5.295]
Public-600, original decoder	1,715 / 1,669 of 1,800	+2.556	[1.389, 3.778]
Public-600, later fallback	1,799 / 1,794 of 1,800	+0.278	[0.000, 0.556]
Hard-110, frozen fallback	207 / 172 of 330	+10.606	[4.848, 16.667]

Table E3. Completed weighting evidence at its original scope. Intervals condition on the fitted pairs. The two new-root panels each contain 768 roots; the final rank-64 models are reused across their subsequent panels. The public fallback reuses original successes. Rows must not be pooled as independent replications, and their effect differences are not isolated capacity effects.

The rank-16 fit differences were −1.172, +8.464 and −1.432 points, with 201 paired wins and 156 losses. Its crossed interval was [−2.865, 7.856]; omitting the strongest fit gave −1.302 points. All six block-level pooled point differences were positive. This earlier result supports consistency of the pooled direction at another operating point, not uniform training-seed robustness. Those models and roots are not merged with the final rank-64 result. The source-native RAP adaptation, using learned-world-model search and final exact verification, scored 66/768. No oracle state repair was available during its search. Its single outcome vector is reused across the three VMGR fits; it does not represent three independent RAP experiments. The unfitted public feedback adaptation scored 14/600, versus 13/600 for its own Direct first attempt. These weak adaptations are diagnostic history rather than competent reproductions of author-reported RAP or feedback performance.

For VALUE−PERMUTED, paired wins/losses are 53/47, 52/60 and 94/27 across the three seeds. Leaving each seed out gives +3.841, +4.753 and −0.130 points. For UNIFORM−PERMUTED the corresponding sensitivity estimates are −1.628, +1.042 and −1.888 points. The mechanism result depends strongly on one trained fit. Neither a negative seed nor a nonpositive interval is removed as an "unlucky model." Additional seeds were not used to search for a positive mean.

Permutation training adds 528 formal updates plus 3 qualification updates, 531 physical updates in total. Formal training takes 24,863.650, 23,472.269 and 24,358.565 seconds across seeds; qualification adds 354.622 seconds. Its 2,304 new inference episodes take 32,961.316 seconds including queue loading and 56,327,904 forward-token positions. These components total 29.447 worker-hours; additional historical work is excluded, not declared zero. The 4,608 existing VALUE/UNIFORM receipts are reused, but their original training and inference costs are not zero.

The historical LEX 163/216 reference remains visible but is nonmatched in recipe and population. Earlier rank 16 results, negative value-aligned variants, capacity searches and the seed-sensitive permutation test form part of the research history. None is pooled into the final public sample or silently relabelled as a new independent confirmation.

F FORMATIVE DIAGNOSTICS AND THE COMPARISON FRAMEWORK

These cases preceded the present VMGR comparison. The candidate procedure was synthesized retrospectively from them, not independently validated by them. We preserve their aggregate conclusions here; they are not additional VMGR benchmarks. The six gates cover source/population validity, semantic object/verifier validity, task/interface competence, optimization/inference resources, construct-valid measurement and claim-bearing recipe provenance. Their number and ordering are not claimed to be minimal or universally sufficient.

Gate	Application and remaining boundary
G1: population	Separate reused development roots, exact-identity-new roots and public tasks. Exact novelty does not certify decontamination.
G2: semantics	Replay executable macros and complete plans with exact transitions. Their action interfaces still differ.
G3: competence	Keep zero-floor and unsuccessful external adaptations diagnostic; they cannot establish superiority over the original methods.
G4: resources	Match support/work and decoder in VALUE–UNIFORM. Report unequal SFT exposure and actual inference work separately.
G5: measurement	Use goal-reaching execution, paired task clusters and explicit failure accounting, not proxy loss or independent-episode counts.
G6: provenance	Distinguish paper, released and locally adapted recipes; retain fixed endpoints, development choices and failed work.

Table F1. Candidate comparison conditions and their application to VMGR. No independent gate-validation claim is made.

Executable code repair: population stop. TRAIN engineering qualification succeeded, but the audited source could not instantiate the declared fresh-problem population. One frozen source census retained 870 eligible rows against 3,712 required candidate positions, a shortfall of 2,842. Bounded text interfaces also failed their qualification conditions. These outcomes limit the feasibility of that population and interface, not the relative merits of generative architectures. No retrospective quota reduction or replacement population is pooled with this case.

Exact GF(2): shared competence floor. Matched 170M system packages reached zero exact passes on all 1,440 paired formal systems. Subsequent new-seed controls did not establish prerequisite task-and-interface competence. A zero paired difference at this floor is not evidence of architectural equivalence. It motivates checking whether the declared interface can support the intended contrast before interpreting a ranking. These are previously reported aggregates; no historical individual panel was reopened for this manuscript integration.

Minimal planning: measurement and provenance. In a local reproduction of the published minimal-planning setup, (Ye et al., 2025) the initial difficulty stratifier measured the shared-node position instead of planning distance. For the six-node generator, the corrected distance is `5 - sep_position`. Unchanged-output replay corrected the six reversed strata. The AR easy-cell condition then passed, while the iterative model remained at zero exact and semantic pass. The original measurement-invalid record remained separate from its corrected diagnostic successor.

The paper specifies 1,200 iterative-model epochs for minimal planning, whereas the audited public path script specifies 100 epochs (Ye et al., 2025; HKUNLP, 2025). The locally executed public recipe produced 1,000 outputs with no exact or semantic successes: 857 were discontinuous and 143 malformed. This does not refute the paper's different training recipe. It limits the conclusion to the executed public-code condition. The code revision is pinned in the reference; the schedule mismatch is not an assertion about every later release.

Decision representation and repair boundaries. Each gate record takes a state from pass, scientific-fail, engineering-invalid, measurement-invalid, not-reached, unavailable and non-applicable. An applicable failed or unavailable prerequisite blocks the corresponding claim. The five terminal labels are admit-package-comparison, unidentified-stop, engineering-invalid, measurement-invalid and exploratory-only. A complete applicable-gate pass supports only the declared package estimand; it does not establish architecture causality. Post-outcome reanalyses of observed evaluations remain exploratory; a successor study must declare its own estimand, frozen decisions and evaluation population. These labels are conditional on the declared contrast, not a universal ranking of studies.

The CPU reference artifact contains a record schema, six synthetic records covering all seven states and five labels, a small GF(2) verifier/estimator fixture and three fixed aggregate stopping replays. Table F2 specifies the checks applied to a supplied label after schema validation. It is not a general dependency-graph executor and does not choose a unique label by universal precedence when several conditions hold.

Supplied terminal label	Implemented required condition
admit-package-comparison	Every applicable state is pass and every applicable gate is predeclared.
unidentified-stop	At least one state is scientific-fail or unavailable.
engineering-invalid	At least one state is engineering-invalid.
measurement-invalid	At least one state is measurement-invalid.
exploratory-only	At least one gate is not predeclared.

Table F2. Record-consistency requirements, not sufficient scientific evidence. All six gate IDs and valid state/count fields are checked before these rules. Non-applicable gates are excluded from admission's predeclaration requirement.

The aggregate replayer returns unidentified-stop for a positive code-source deficit, either GF(2) arm at zero exact success, or a planning recipe mismatch or nonpositive exact-success control. Clearing an individual prerequisite returns no terminal admission; a full record is still required. Isolated standard-library execution passes 16 unit tests, including malformed records and partial-pass boundaries. Synthetic records are not empirical admitted comparisons. These checks do not establish provenance truth or causal assumptions. The artifact exists as a local review candidate, not an already published licensed release.

Connection to VMGR. The formative cases establish why source, semantics, competence, measurement and recipe identity need explicit treatment. A later VMGR-specific diagnosis found 5,028 value-aligned and 5,592 lexical root–prompt pairs, with neither support containing the other (Section 1). Shared-support weighting addresses that narrower attribution problem. It does not repair the code population, establish GF(2) competence or reproduce the missing iterative recipe. VMGR's positive results therefore extend the research program without erasing its unresolved cases.

F.1 Worked claim-specific applications

These examples apply the candidate conditions retrospectively to completed VMGR experiments. They explain allowed and blocked conclusions; they do not create predeclared admission records or independently validate the framework. The experimental protocols retain their original prospective or post-hoc status.

Claim A: target weighting helps these fixed fits on the final new-root panel. The estimand is the block-balanced paired success difference between VALUE and UNIFORM, with support, optimization rules and decoder fixed.

Gate	Evidence and claim-specific implication
G1	768 exact-identity-new roots, fixed selection and exclusions. Admits the declared panel, not isomorphism novelty or pretraining cleanliness.
G2	Same generated macro grammar, transition function and exact-goal endpoint. No interface change explains this contrast.
G3	Both systems exceed 80% success. The contrast is not at a common floor; no minimum positive effect is required for interpretability.
G4	Same support, token exposure, schedule and decoder rules. Actual inference work is measured and can differ with stopping; equal FLOPs are not claimed.
G5	Complete paired records and task-cluster intervals yield +3.646 pp [2.040, 5.295]. This is conditional on three fitted pairs.
G6	Fixed rank-64, step-176 checkpoints and locally frozen analysis precede outcomes. This does not restore a randomized architecture contrast.

Table F3. Application to the identifying weighting contrast. The narrow fixed-fit claim is supported. General improvement over new training seeds, other domains, architectures or all strong methods remains unsupported.

Claim B: the executed VALUE package has higher Hard-110 success than the executed best-first SFT package. The estimand is a task-weighted difference between these specified systems, not an isolated training-objective effect.

Gate	Evidence and claim-specific implication
G1	All 110 source hard tasks, with fixed models and settings before original outcomes. Admits this configuration, not every planning benchmark.
G2	Same exact execution endpoint, but macro and primitive-action proposal interfaces differ. Package ranking is possible; representation attribution is not isolated.
G3	SFT solves 13/330 and was competent on smaller tasks. Its poor transfer is observed, not proof that SFT or search generally fails.
G4	Training exposure and actual inference work differ and are reported. Success-first comparison is allowed; equal-resource superiority is not established.
G5	110 paired task clusters give 207/330 versus 13/330 and +58.788 pp. Reused fits are not new training replications.
G6	This is the implemented best-first adaptation, not an author-faithful ToT/RAP or universal external-SOTA baseline.

Table F4. Application to an unequal-resource system contrast. The measured package advantage is supported. Weighting attribution comes from the separate VALUE–UNIFORM comparison, not from this larger SFT gap.

A failed prerequisite blocks a dependent inference, not every observation. The GF(2) case can still report both packages' zero success even though it cannot infer architecture equivalence. Likewise, a low-performing Hard-110 baseline remains a valid descriptive system result without becoming a competent stand-in for an entire method family. This distinction makes the procedure useful without requiring equal budgets or positive effects for all comparisons. Neither worked example establishes the gates' sufficiency.