WHAT IS ONE SEARCH OPERATION WORTH? EVALUATING ITS DOWNSTREAM VALUE IN AUTOMATIC AGENT DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic agent design searches for better prompts, agent connections, and workflows. Yet a strong final system does not reveal which individual search decisions improved it. We introduce a protocol for evaluating a search operation through the optimization that follows it. From the same recorded state, we execute the original operation or a specified alternative, then resume optimization under the same objective and budget. We apply this protocol to GPTSwarm, AFlow, MaAS, and EvoMAS across six benchmarks. A dedicated GPTSwarm–MMLU study includes 480 executions at 20 states, with same-operation controls and separate runs for selection and validation. Operations selected using four repetitions per alternative show an apparent mean advantage of 2.17 percentage points over the unselected alternatives. Four additional repetitions on the same evaluation items estimate this advantage at 0.65 points (95% CI $[-0.56, 1.82]$), yielding a paired selection–validation gap of 1.52 points $[0.41, 2.70]$. GPTSwarm and EvoMAS/MMLU also show measurable separation in subsequent search trajectories, while estimated associations with downstream value remain imprecise. These measurements distinguish how an operation changes search from how well a selected operation performs on new executions, enabling evaluation of individual decisions within automatic agent design.

1 INTRODUCTION

Automatic agent design moves system development into a search loop. Here, *search* refers to the iterative generation, evaluation, and revision of candidate agent systems. A *search operation* is a specified choice or update within this outer optimization loop. An optimizer can revise prompts, connect agents, expand a workflow, or select parents for new systems (Khattab et al., 2024; Zhuge et al., 2024; Hu et al., 2025; Zhang et al., 2025b; Hu et al., 2026). Each decision spends computation and changes the candidates available to later search. Understanding which decisions are worth making is therefore central to improving the optimizers themselves.

Consider a workflow optimizer choosing between two parent programs. Expanding one parent produces a promising child. Was that parent a good choice? A different parent could have produced an equally strong child; repeating the same choice could have produced a weaker one; and subsequent search could have changed their ordering. The final system score combines all of these events. To evaluate the choice, we need to return to the state where it was made, execute an alternative, and follow both branches through the optimization that comes next.

This example exposes three distinct questions. *Does changing the operation steer search? Does the resulting outcome difference exceed the variation seen when repeating the same operation? Does an operation selected from limited observations retain an advantage in independent runs?* Trajectory distances relative to within-operation variation address the first question; endpoint contrasts with same-operation controls address the second; new executions of a previously fixed choice address the third. Together, these measurements help diagnose search, choose evaluation budgets, and assess proposed guidance signals.

We develop a factual-state evaluation protocol around these questions (Figure 1). We restore a recorded optimizer state, replace one executed operation with a prespecified admissible alternative,

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.

and resume the native optimizer with the same objective, rule, and budget. Later updates and routing remain free to diverge. Repeated endpoints estimate the operation's value relative to that alternative. Same-operation branches provide an execution reference, and separate selection and validation waves evaluate whether a finite-budget choice transfers to new runs. We implement the protocol at graph-update, workflow-parent, controller-draw, and evolutionary-parent boundaries in GPTSwarm, AFlow, MaAS, and EvoMAS.

The resulting evidence connects process changes to the reliability of decisions. In a new GPTSwarm–MMLU study, all 480 branches complete across 20 frozen states. Choosing an operation from four repetitions gives an apparent advantage over the unselected alternative of 2.17 percentage points (pp); four new repetitions give 0.65 pp, with a 95% interval of $[-0.56, 1.82]$. The positive paired gap, 1.52 pp $[0.41, 2.70]$, documents a reduction in apparent advantage; the validation interval leaves the mean benefit uncertain. The intervention also separates subsequent graph trajectories. EvoMAS/MMLU shows separation between subsequent offspring configurations; BBEH-mini provides a separate full-trajectory diagnostic. These results motivate measuring search steering and independently evaluated value together.

Our contributions are threefold:

- **An operation-level evaluation protocol** that combines factual-state interventions, native optimizer continuation, and explicit endpoints across four optimization mechanisms.
- **An independent choice-evaluation study** with same-operation controls, quantifying the gap between the apparent and independently evaluated advantage of finite-budget operation selection in GPTSwarm.
- **A cross-mechanism empirical analysis** that separates trajectory changes from downstream value, and characterizes how local signals, score resolution, and continuation affect the evidence for an operation.

2 Background and Related Work

2.1 Automatic search over agent systems

An agent system combines language-model calls, prompts, tools, and communication or control flow. Automatic design methods optimize different parts of this object. DSPy compiles modular language-model programs against a metric (Khattab et al., 2024); ADAS searches executable agent programs (Hu et al., 2025); and AgentSquare searches modular designs (Shang et al., 2025). GPTSwarm updates communication graphs, AFlow expands code-represented workflows, MaAS samples query-conditioned architectures, and EvoMAS evolves multi-agent configurations (Zhuge et al., 2024; Zhang et al., 2025b;a; Hu et al., 2026). These mechanisms expose different choices: a utility binding, a parent, a controller draw, or a parent tuple. Each choice is an outer-loop search operation whose consequences include the candidates and updates that follow.

Evaluating the returned system establishes what the full search procedure achieved under its budget. Evaluating an individual operation additionally requires specifying where the comparison starts and how search continues. For example, a strong workflow need not be the best parent to expand, and a large graph update need not improve later task performance. The native objective also differs from the measurement endpoint: an optimizer may track the best system so far while an experiment measures the next generated child. We retain these distinctions when instantiating a common evaluation protocol.

2.2 Evaluating choices inside adaptive search

Value functions and metareasoning formalize the consequences of an action under a continuation policy (Sutton and Barto, 2018; Russell and Wefald, 1991; Hay et al., 2012). Neural architecture search distinguishes search-policy quality from the quality of the selected architecture (Elsken et al., 2019; Ying et al., 2019; Yu et al., 2020), and component-contribution methods investigate whether internal search parameters reflect useful choices (Wang et al., 2021; Xiao et al., 2022). These ideas motivate state-matched comparisons, explicit continuation, and uncertainty for the estimated difference. Our target is a recorded outer-loop operation followed by further optimizer adaptation.

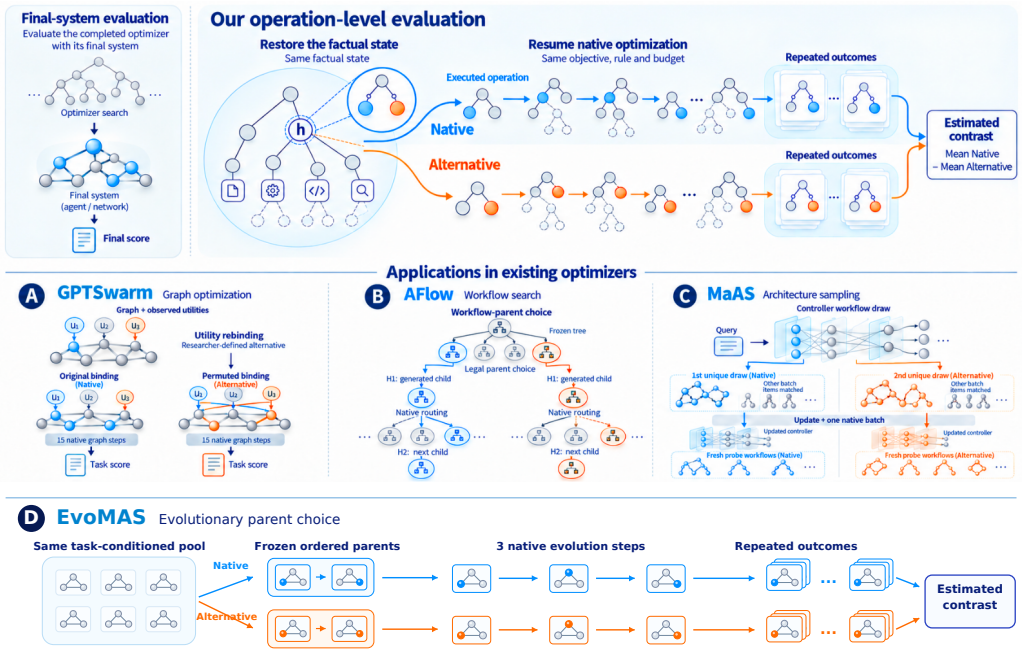


Figure 1: **From final-system evaluation to operation-level evaluation.** Restore a factual state, replace one operation, and resume native optimization before comparing repeated outcomes. A–C show utility rebinding, workflow-parent choice, and controller draws. D adds EvoMAS: freeze an ordered parent tuple or its alternative, then continue three native evolution steps. Blue and orange denote Native and alternative branches. The diagrams illustrate the protocol; independent validation appears in Figure 2.

Approach	Focal unit	Downstream comparison
HGM (Wang et al., 2026)	Ancestor selected for expansion	Expected final utility (Q-value); clade-based estimates
SHOR (Ong et al., 2026)	Component-priority judgement	Subsequent harness optimization
TreeCredit (Liu et al., 2026b)	Reasoning operator	Completed reasoning suffix
AutoML Evaluation Agent (Du et al., 2026)	Pipeline decision	Pipeline outcomes
C3 (Chen et al., 2026); CAR (Shah, 2026)	Message or agent step	Frozen or same stochastic policy
This study	Outer-loop search operation	Native updates, routing and candidate generation

Table 1: **Evaluation units and downstream comparisons.** The approaches measure complementary objects in adaptive systems. Our comparison starts from a recorded state and retains the original optimizer's updates after the focal intervention.

HGM provides a direct conceptual precedent (Table 1). Its global metaproductivity formalizes the expected final-agent utility after selecting a parent and continuing search, explicitly corresponding to a Q-value; HGM guides expansion using estimated clade-level outcomes (Wang et al., 2026). SHOR evaluates component-priority rankings against subsequent harness optimization (Ong et al., 2026). TreeCredit compares reasoning operators from shared states using completed suffixes (Liu et al., 2026b). AutoML decision evaluation and counterfactual agent replay also examine intermediate choices (Du et al., 2026; Chen et al., 2026; Shah, 2026). We operationalize state-matched intervention comparisons across four search mechanisms, combining native continuation, same-operation controls, and separate selection and evaluation executions.

Repeated evaluation is also established: RAAS uses repeated, matched-query trials for architecture comparison (Yang et al., 2026). Building on action-value evaluation and held-out validation, our contribution is an executable operation-level protocol and its controlled empirical assessment across optimizer boundaries. Fixing a choice in one wave and evaluating it in another links estimation precision to a concrete decision. Local signals and predictive guidance (Li et al., 2026; Yuan et al., 2026) can use these independent downstream outcomes as evaluation targets.

3 EVALUATING A SEARCH OPERATION

3.1 RESTORE, INTERVENE, AND CONTINUE

Let h be a recorded optimizer decision state, d^N the native operation, and d^{CF} an alternative fixed before its downstream outcomes. The intervention boundary determines the required state: graph and optimizer parameters, a workflow tree, a controller, or a task-conditioned evolutionary pool. We restore this state separately for each branch, execute the assigned operation, and resume the original optimizer with the same objective, continuation rule C_H, and budget H. Subsequent parameters, candidates, and routes may differ; their divergence is part of the operation's consequence.

Alternatives must be executable at the declared boundary. GPTSwarm uses a researcher-defined rebinding of already observed utilities; AFlow uses another legal parent; MaAS uses another controller draw; EvoMAS uses a frozen alternative ordered parent tuple. The alternative-selection rule and endpoint specify the comparison. Appendix A gives the implementation details and integrity checks.

3.2 DOWNSTREAM VALUE AND PROCESS SEPARATION

For endpoint J_H, define the conditional value and Native-relative contrast

$$V_H(h,d) = \mathbb{E}[J_H \mid h, \mathrm{do}(d), C_H], \qquad A_H(h) = V_H(h,d^N) - V_H(h,d^{CF}). \tag{1}$$

The expectation covers subsequent search and evaluation randomness. Positive A_H favors the executed operation. Averaging valid branch outcomes over repeated continuations gives

$$\widehat{A}_H(h) = \frac{1}{|\mathcal{R}_h|} \sum_{r \in \mathcal{R}_h} \left(J^N_{H,r} - J^{CF}_{H,r}\right), \tag{2}$$

where $\mathcal{R}_h$ is the common set of valid repetition IDs. With incomplete branches, this is a completed-support estimate; planned-cohort missing-outcome bounds are reported separately. A newly generated child's score, a held-out task score, and the native cost-sensitive reward each define a different value question.

To measure whether operations change the search process, let D be a declared trajectory distance. Prespecified measures and retrospective sensitivities are identified separately. With R repetitions in each arm, define

$$B = \frac{1}{R^2} \sum_{r,s} D(\tau^N_r, \tau^{CF}_s), \quad W = \frac{1}{2} \sum_{b \in \{N,CF\}} \frac{1}{\binom{R}{2}} \sum_{r<s} D(\tau^b_r, \tau^b_s), \quad S = B - W. \tag{3}$$

S compares cross-operation distance with within-operation variation. GPTSwarm uses subsequent edge-probability vectors; EvoMAS uses canonical configuration fields. Continuation-only analyses exclude the intervention-stage representation; full-trajectory measures are labeled separately. Each distance is interpreted within its framework.

3.3 SAME-OPERATION CONTROLS AND INDEPENDENT VALIDATION

Native-A and Native-B apply the same frozen operation with independent executions; CF applies the fixed alternative. In GPTSwarm, contrasts with a common Native-B reference compare the magnitude of Native–Native and Native–CF endpoint differences. Their signed values are also retained. This control measures variation under repeated execution of the same operation and reveals arm offsets.

At each state, wave 1 selects the operation with the larger estimated outcome. Wave 2 evaluates that fixed choice against the unselected alternative using disjoint executions. Both use the same frozen

Table 2: **Study overview and evidence roles.** N counts states or runs; R is repetitions per arm. Slashes show planned/analyzed support. HumanEval ordering uses 14/21 states; EvoMAS/MMLU has 20 complete task endpoints. Fresh GPTSwarm revisits existing states. AFlow's later three-arm waves appear in Table 3.

Setting	Focal comparison	Continuation / endpoint	N	R	Role
GPTSwarm MMLU	Native utility binding vs. nonidentity rebinding	15 native steps; held-out task score	30	4	Historical signals
GPQA-Diamond	Same rebinding	15 native steps; task score	30	4	Historical signals
MMLU, fresh	Native-A / Native-B / CF	Same 153-item endpoint; two waves of four	20	8	Primary validation
AFlow HumanEval	Legal parent choices	Child score at H_1 and after +1 native step (H_2)	21	≤ 3	Historical diagnostics
MBPP	All committed legal parents	Generated child score	24	3	Historical diagnostics
GSM8K	Native vs. uniform over the same parents	Weighted generated-child score	20/17	≤ 3	Historical allocation
MaAS HumanEval	First vs. second unique controller draw	Focal update +1 native batch; fresh probe utility	29	3	Historical signals
EvoMAS MMLU	Frozen ordered parent tuple vs. alternative	3 evolution steps; task and reward endpoints	30/24	3	Process / sensitivity
BBEH-mini	Same parent intervention	3 steps; retained reward/state table	20/19	3	Process diagnostic

states and evaluation items. Let $\widehat{A}_i^{(1)}$ and $\widehat{A}_i^{(2)}$ be their Native-A–CF contrasts, and $s_i = \operatorname{sign}(\widehat{A}_i^{(1)})$. We evaluate

$$G_{\text{select}} = \frac{1}{N}\sum_i |\widehat{A}_i^{(1)}|, \qquad G_{\text{validate}} = \frac{1}{N}\sum_i s_i \widehat{A}_i^{(2)}. \tag{4}$$

Both quantities compare the selected operation with the unselected alternative; an exact selection tie contributes zero. Choosing the larger first-wave estimate can make $G_{\text{select}} > 0$ even for equally valuable operations. The paired gap records how this advantage changes in new executions on the same states and endpoint items. Generalization to new states or questions is a separate target.

4 Experimental Setup

Table 2 summarizes the evidence. We analyze each setting in its own units and retain the acquisition order. The new GPTSwarm study revisits 20 existing MMLU states; its 480 branches are repeated observations of those states. Historical GPTSwarm includes 30 MMLU and 30 GPQA-Diamond states. AFlow covers HumanEval, MBPP, and GSM8K; EvoMAS adds BBEH-mini (Hendrycks et al., 2021; Rein et al., 2024; Chen et al., 2021; Austin et al., 2021; Cobbe et al., 2021; Kazemi et al., 2025).

New GPTSwarm validation. Each state runs Native-A, Native-B, and CF eight times, in two waves of four. After the frozen utility-binding intervention, branches continue for 15 native updates and are scored on the same 153 MMLU items. These items have no exact-prompt overlap with the 285 training items. Arm-specific local seeds and new API requests separate executions; provider-side deterministic seeding is unavailable. The primary endpoint is G_{validate}. The trajectory analysis was specified before wave 2 and before inspecting fresh outcome summaries. All 480 planned branches complete. Appendix B reports the frozen design, controls, and costs.

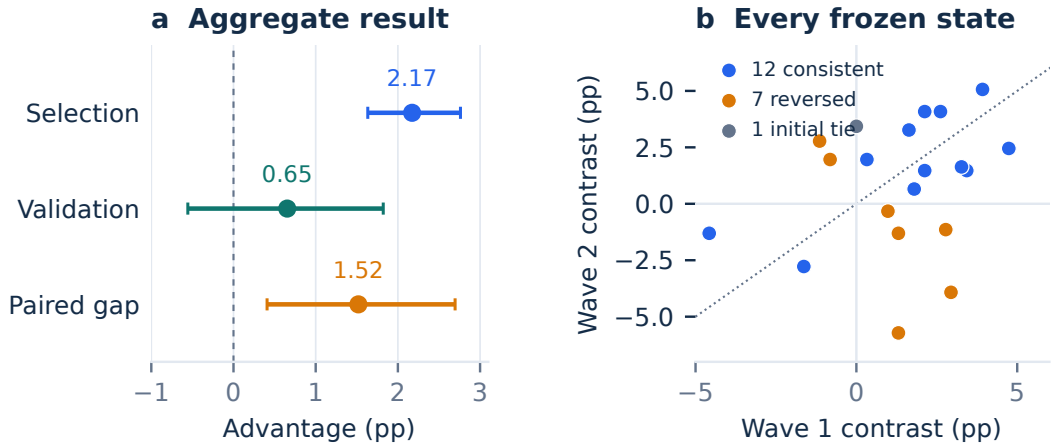


Figure 2: **Evaluating a choice on new executions.** Left: selection advantage, independent validation advantage, and their paired difference, in accuracy percentage points. Bars are 95% state-bootstrap intervals. Right: all 20 states; positions show signed Native-A–CF contrasts in the two disjoint four-repetition waves. Color identifies preserved direction, reversal, or a first-wave tie. The diagonal marks equal contrasts. Both waves use the same 153 endpoint items.

EvoMAS and earlier cohorts. EvoMAS freezes the selected ordered pair of parents from a common six-configuration pool and compares it with a prespecified alternative, then runs three evolution steps. Its unit is a task-conditioned decision history, rather than an independently evolved population. On MMLU, the frozen primary task endpoint is exact-match correctness; the native judge/resource reward is secondary. BBEH-mini supports a separate analysis of the frozen state table (Appendix C). AFlow retains its original ordering and horizon studies and two later six-repetition three-arm waves. Historical MaAS probes fresh workflows after controller updates with $J = \text{score} - 3\,\text{cost}$. Appendices D and E report these settings.

Inference and resources. The main model route is `gpt-4o-mini-2024-07-18`; Qwen2.5-72B supplies secondary historical evidence. Intervals resample complete state records, retaining their nested repetitions, with 10,000 draws for new GPTSwarm and EvoMAS analyses. Evaluation items remain fixed. The GPTSwarm primary outcome is distinguished from unadjusted secondary comparisons; EvoMAS offspring-only distances are retrospective sensitivities. Completion accounting and finite missing-outcome bounds accompany incomplete cohorts. Appendix F reports resources and acquisition records.

5 Results

5.1 Independent choice evaluation and same-operation controls

In GPTSwarm, choosing between Native-A and CF using the first wave produces an apparent advantage of 2.17 pp $[1.63, 2.76]$. The same fixed choices achieve 0.65 pp $[-0.56, 1.82]$ in the second wave. Their secondary paired gap is 1.52 pp $[0.41, 2.70]$ (Figure 2). Twelve states preserve the selected direction, seven reverse it, and one has an exact first-wave tie. The primary validation interval includes both a modest loss and a gain, while the paired gap quantifies the reduction from the selection-stage estimate.

Same-operation executions put these differences in context. With eight-repetition arm means and a common Native-B reference, the mean absolute Native–Native and Native–CF differences are 1.97 and 1.99 pp. Their paired difference is 0.02 pp $[-0.89, 0.95]$. More directly, independently evaluated cross-operation choices have -0.25 pp $[-2.00, 1.50]$ advantage relative to the analogous same-operation choices. The two-wave study quantifies the observed selection–validation gap, while these controls leave its operation-specific component unresolved.

The signed Native-B–Native-A mean is -1.00 pp $[-1.96, -0.06]$, an unadjusted offset despite identical operations. Within-wave slots were shuffled before acquisition. The design does not isolate finite sampling, timing, and provider contributions to the offset; results describe the recorded execution regime. Appendix B.2 reports all signed contrasts and selection controls.

5.2 Operations separate subsequent search trajectories

Changing the utility binding in GPTSwarm produces separation that persists through subsequent graph updates (Appendix C, Figure 7). After excluding the immediate post-intervention state, the mean Native-A–CF trajectory separation is $S = 0.00628$ with a 95% interval of $[0.00427, 0.00825]$; 18 of 20 state estimates are positive. The same-operation Native-A–Native-B reference is -0.000008 $[-0.000753, 0.000889]$. Their paired difference is 0.00629 $[0.00404, 0.00849]$, and each four-repetition wave separately has positive mean cross-operation separation. Thus, the changed operation leaves a measurable signature in the continuing optimizer.

EvoMAS extends the comparison to generated configurations. Its prespecified MMLU endpoint is task accuracy: the finite three-repetition contrast is zero in all 20 states with complete task endpoints, making the empirical separation–accuracy correlation undefined. Each state evaluates one task per repetition: 15 states are always correct, four always incorrect, and one scores $2/3$ in both arms. We report secondary native reward separately below.

The retrospective MMLU offspring-only distance excludes adapted parents and operation labels. Separation is positive in 23 of 24 complete states, with mean 0.296 $[0.235, 0.359]$. Allowing six incomplete states arbitrary separation in $[-1, 1]$ gives planned-cohort mean bounds $[0.0368, 0.4368]$, separately from sampling uncertainty. BBEH-mini's frozen full-trajectory measure is positive in all 19 valid states and all 14 after prespecified repair exclusions. It includes adapted-parent and action components; its retained state table cannot support the same offspring-only reconstruction.

GPTSwarm wave-1 separation has $\rho = 0.00$ $[-0.45, 0.47]$ with the independent wave-2 accuracy contrast. EvoMAS/MMLU offspring separation has $\rho = 0.032$ $[-0.418, 0.469]$ with secondary native reward; BBEH's full-trajectory association is $\rho = 0.302$ $[-0.187, 0.687]$. These compare unsigned separation with Native-relative signed contrasts, and leave association direction and size imprecise. Positive process separation thus motivates measuring downstream value separately.

Mean EvoMAS reward contrasts are -0.114 $[-3.547, 3.060]$ on MMLU and -3.008 $[-10.922, 3.830]$ on BBEH. Reward combines a judge score with token/time penalties; in 15 of 24 complete MMLU states, all observed reward difference comes from the resource term. Appendix C separates task, judge, and resource measurements.

5.3 Local signals against independent downstream outcomes

We assess local signals against new executions of the downstream comparison. On the 20 revisited GPTSwarm/MMLU states, historical edge-probability displacement has $\rho = 0.13$ $[-0.35, 0.56]$ with the fresh wave-2 contrast (Figure 3). The immediate score gap has $\rho = 0.08$ $[-0.47, 0.57]$, and signed expected-edge-count change has $\rho = -0.17$ $[-0.67, 0.35]$. Historical Native terminal performance has a positive secondary association, $\rho = 0.61$ $[0.16, 0.86]$, but requires a downstream evaluation. Table 5 retains all declared signals. The fresh estimates provide limited precision for using inexpensive signals to guide operation choice.

The historical cohorts supply context: GPTSwarm/MMLU has $\rho = 0.60$ $[0.29, 0.80]$ ($N = 30$), and secondary Qwen evidence has $\rho = 0.38$ $[0.03, 0.65]$. Different sample sets and executions prevent interpreting these coefficients alone as a significant change in association. In MaAS, controller preference has $\rho = 0.29$ $[-0.11, 0.65]$ with cost-sensitive utility across 29 runs, with mean contrast 0.00560 $[-0.01188, 0.02583]$. These results distinguish a historical diagnostic association from evidence that a signal anticipates new outcomes.

5.4 Endpoint, continuation, and completion sensitivity

Historical AFlow illustrates endpoint resolution under limited repetition. Among tied-guidance states, branch means differ in 10/14 HumanEval and 7/24 MBPP states; nine and five of these differences arise entirely from one of three repetitions (Figure 9, Appendix E). Scores concentrate at 0.8 (69/84

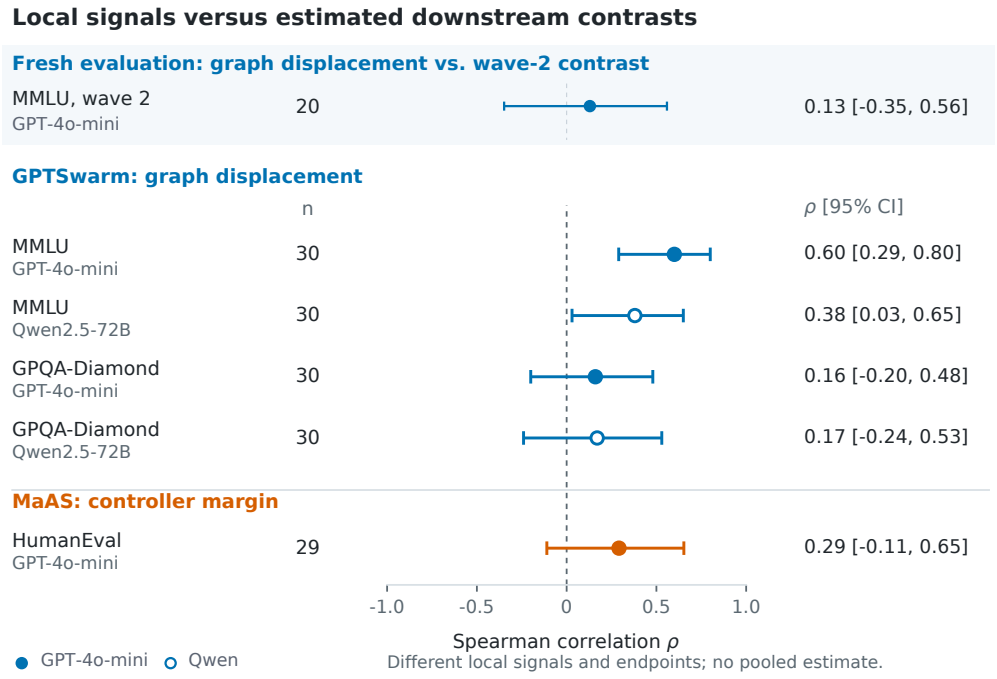


Figure 3: **Local signals against fresh and historical outcomes.** The added top row relates historical graph displacement to the fresh wave-2 contrast at 20 states. The original rows retain historical GPTSwarm ($N = 30$ each) and MaAS ($N = 29$) associations. Bars are 95% intervals; Qwen is secondary and its MMLU contrasts are reconstructed. Different samples and endpoints are analyzed separately.

HumanEval outcomes) and 0.7 (134/144 MBPP outcomes); none reaches 1.0. This is concentration on a coarse ten-item grid. Fresh GPTSwarm uses a finer 153-item endpoint, has no perfect or zero-scoring branch, and still exhibits the selection–validation gap.

AFlow's next native expansion can select an earlier tree node, rather than expand the immediate child: 83 of 117 completed historical routes do so. On these 21 states, the mean absolute Native–CF estimate changes from 0.03889 at the immediate child to 0.01429 at the next generated child. The paired change is -0.02460 $[-0.05397, -0.00476]$; intersecting valid IDs across horizons gives -0.02302 $[-0.05238, -0.00317]$. Single-state deletion retains a negative mean change, while the median state change is zero.

Table 3 retains both new AFlow waves. These show why the comparison includes its execution regime and support. Under the current scorer/runtime, the first six-repetition wave gives a shared-support change of 0.1263 $[0.0451, 0.2228]$, the added independent six-repetition wave gives -0.0180 $[-0.0530, 0.0140]$, and the secondary pooled description gives 0.0404 $[0.0076, 0.0775]$. The historical Native and CF completion counts are 57/63 and 60/63; bounding missing scores in $[0, 1]$ gives a planned historical change range of $[-0.07302, 0.08095]$. Changes in these finite absolute estimates reflect endpoint, support, and execution variability as well as possible value changes. Appendix E reports both new waves, same-parent controls, additional-item checks, and missingness.

Table 3: **Independent AFlow repetition waves at 21 states.** H_1/H_2: mean absolute contrasts on horizon-specific support. Shared-ID changes need not equal $H_2 - H_1$. Units: score proportions.

Comparison	H_1	H_2	Shared-ID change	95% CI
Initial six repetitions	Shared valid repetitions per state: 1–6			
Native-A vs. CF	0.0140	0.1440	$+0.1263$	$[0.0451, 0.2228]$
Native-A vs. Native-B	0.0220	0.1180	$+0.1020$	$[0.0299, 0.1871]$
Added independent six repetitions	Shared valid repetitions per state: 2–6			
Native-A vs. CF	0.0387	0.0415	-0.0180	$[-0.0530, 0.0140]$
Native-A vs. Native-B	0.0450	0.0160	-0.0387	$[-0.0873, -0.0010]$
Combined twelve repetitions (secondary)	Shared valid repetitions per state: 3–12			
Native-A vs. CF	0.0231	0.0775	$+0.0404$	$[0.0076, 0.0775]$
Native-A vs. Native-B	0.0283	0.0635	$+0.0331$	$[-0.0151, 0.0825]$

6 Discussion

A high-scoring candidate does not by itself show that the search decision producing it was valuable, because an adaptive optimizer can redirect effort after that decision. HGM formalizes this distinction through future-search utility (Wang et al., 2026). AFlow illustrates it empirically: in 83 of 117 completed routes, the next expansion selects another existing node rather than the immediate child. A parent choice can therefore matter through the search it sets in motion, suggesting that it should be evaluated through its continuation rather than only its immediate score.

Following the continuation, however, does not resolve whether the resulting change in search is useful. GPTSwarm and EvoMAS/MMLU show trajectory separation beyond within-operation variation, yet associations between that separation and downstream value remain imprecise. Structural change therefore establishes that an operation redirects search, but not by itself that the redirection is beneficial. Task and cost endpoints address that separate question. The state, operation, alternative, continuation budget, and endpoint jointly define the value question.

The same distinction becomes important when local signals guide search. In GPTSwarm, historical graph displacement has an imprecise association with fresh downstream outcomes. Separately, score-based choices show a 2.17 pp selection advantage but only a 0.65 pp validation advantage (95% CI $[-0.56, 1.82]$). Together, these results suggest evaluating both whether a signal is informative and whether the choices it induces retain value on independent continuations. Same-operation runs provide an execution reference for interpreting such comparisons.

7 Limitations

Independent validation focuses on 20 recorded GPTSwarm states, one model route, and 153 fixed evaluation items, characterizing repeatability in the recorded setting. Newly sampled states, additional routes, and unseen items would broaden the scope of this evidence. The -1.00 pp Native-B–Native-A offset provides an additional reference for execution-level differences; its source merits further controlled study. The complementary cohorts extend mechanism coverage under different endpoint and logging regimes (Appendices C, E, and F). Longer horizons and closed-loop evaluation of independently validated signals are natural extensions of the present study. These considerations leave the central methodological message unchanged: search steering, downstream value, and choice reliability require distinct evidence. These limitations mainly concern the breadth of the empirical evidence, while the central methodological distinctions are supported within the settings studied here.

8 Conclusion

Our protocol evaluates search decisions through the optimization that follows them. Explicit interventions, native continuation, same-operation controls, and independent validation distinguish evidence about search steering, downstream value, and choice reliability. For automatic agent design, this framing asks not only whether a search decision changes subsequent search, but whether its apparent advantage persists through continued optimization and independent evaluation.

Reproducibility Statement

The accompanying source package contains machine-readable result tables, plotting code, and provenance hashes. Appendices A–F specify interventions, estimators, completion rules, and execution records. Historical calculations and new measurements are kept separate. Large raw response archives are indexed in the research checkpoint records.

AI Use Statement

AI is used for manuscript proofreading. Figures were generated using AI under the guidance of the authors.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program Synthesis with Large Language Models, 2021. URL https://arxiv.org/abs/2108.07732.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating Large Language Models Trained on Code, 2021. URL https://arxiv.org/abs/2107.03374. arXiv:2107.03374.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact Is Easier: Credit Assignment for Cooperative LLM Agents, 2026. URL https://arxiv.org/abs/2603.06859v2. Version 2.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, 2021. URL https://arxiv.org/abs/2110.14168.

Gaoyuan Du, Amit Ahlawat, Xiaoyang Liu, and Jing Wu. A Framework for Assessing AI Agent Decisions and Outcomes in AutoML Pipelines, 2026. URL https://arxiv.org/abs/2602.22442.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural Architecture Search: A Survey. *Journal of Machine Learning Research*, 20(55):1–21, 2019. URL https://www.jmlr.org/papers/v20/18-598.html.

Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. Selecting Computations: Theory and Applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 346–355, 2012. arXiv:1207.5879.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2021.

Shengran Hu, Cong Lu, and Jeff Clune. Automated Design of Agentic Systems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=t9U3LW7JVX.

Yuntong Hu, Yuting Zhang, Matthew Trager, Yi Zhang, Shuo Yang, Wei Xia, and Stefano Soatto. EvoMAS: Evolutionary Generation of Multi-Agent Systems. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.06511v4.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, et al. BIG-Bench Extra Hard. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://aclanthology.org/2025.acl-long.1285/.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f1cf02ce09757f57c3b93c0db83181e0-Abstract-Conference.html.

Yu Li, Lehui Li, Zhihao Wu, Qingmin Liao, Jianye Hao, Kun Shao, and Fengli Xu. AgentSwift: Efficient LLM Agent Design via Value-Guided Hierarchical Search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):31843–31851, 2026.

Yiqing Liu, Zihao Wang, Hantao Yao, Wu Liu, and Yongdong Zhang. Shared Prefixes, Better Credit: Adaptive Routing for Multi-Agent Reasoning, 2026b. URL https://arxiv.org/abs/2608.02291.

Kai Tzu-iunn Ong, Minseok Kang, Dongwook Choi, Junhee Cho, Seungju Kim, Seungwon Lim, Geunha Jang, Minwoo Oh, Bogyung Jeong, Sunghwan Kim, Taeyoon Kwon, and Jinyoung Yeo. Towards Direct Evaluation of Harness Optimizers via Priority Ranking, 2026. URL https://arxiv.org/abs/2605.22505.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling*, 2024.

Stuart Russell and Eric Wefald. Principles of Metareasoning. *Artificial Intelligence*, 49(1–3):361–395, 1991. doi: 10.1016/0004-3702(91)90015-C.

Jaineet Shah. Causal Agent Replay: Counterfactual Attribution for LLM-Agent Failures, 2026. URL https://arxiv.org/abs/2606.08275v1.

Yu Shang, Yu Li, Keyu Zhao, Likai Ma, Jiahe Liu, Fengli Xu, and Yong Li. AgentSquare: Automatic LLM Agent Search in Modular Design Space. In *International Conference on Learning Representations*, 2025.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

Wenyi Wang, Piotr Piękos, Li Nanbo, Firas Laakom, Yimeng Chen, Mateusz Ostaszewski, Mingchen Zhuge, and Jürgen Schmidhuber. Huxley-Gödel Machine: Human-Level Coding Agent Development by an Approximation of the Optimal Self-Improving Machine. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=T0EiEuhOOL.

Ruochen Wang, Minhao Cheng, Xiangning Chen, Xiaocheng Tang, and Cho-Jui Hsieh. Rethinking Architecture Selection in Differentiable NAS. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2108.04392.

Han Xiao, Ziwei Wang, Zheng Zhu, Jie Zhou, and Jiwen Lu. Shapley-NAS: Discovering Operation Contribution for Neural Architecture Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2206.09811.

Jiayi Yang, Guancheng Wan, Man Zhang, and Mang Ye. RAAS: LLM Agentic System Architecture Search with GRPO. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 34470–34479, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_RAAS_LLM_Agentic_System_Architecture_Search_with_GRPO_CVPR_2026_paper.html.

Chris Ying, Aaron Klein, Eric Christiansen, Esteban Real, Kevin Murphy, and Frank Hutter. NAS-Bench-101: Towards Reproducible Neural Architecture Search. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 7105–7114, 2019. URL https://proceedings.mlr.press/v97/ying19a.html.

Kaicheng Yu, Christian Sciuto, Martin Jaggi, Claudiu Musat, and Mathieu Salzmann. Evaluating the Search Phase of Neural Architecture Search. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/pdf?id=H1loF2NFwr.

Lingzhi Yuan, Chenghao Deng, Fangxu Yu, Souradip Chakraborty, Mohammad Rostami, and Furong Huang. FlowBank: Query-Adaptive Agentic Workflows Optimization through Precompute-and-Reuse, 2026. URL https://arxiv.org/abs/2606.11290.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. Multi-agent Architecture Search via Agentic Supernet. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 75834–75852, 2025a.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating Agentic Workflow Generation. In *International Conference on Learning Representations*, 2025b.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language Agents as Optimizable Graphs. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62743–62767, 2024.

A PROTOCOL AND IMPLEMENTATION DETAILS

Appendix guide. This section defines the execution contract and framework boundaries. Appendix B contains the GPTSwarm primary result and complete controls; Appendix C contains EvoMAS endpoint and distance audits; Appendix D retains historical local-signal evidence; Appendix E reports AFlow repetition, horizon, and missingness results; and Appendix F indexes reproducibility and resource records.

A.1 EXECUTION CONTRACT

For each eligible state, the protocol performs the following steps.

1. Record the factual state, executed operation, admissible alternatives, and the rule selecting the comparator. Fix the endpoint, native continuation budget, repetition schedule, and failure treatment.
2. Restore an isolated copy for each scheduled branch and verify the state and operation hashes. Retain the observed utilities where the boundary is after scoring.
3. Execute the assigned operation once. Resume the original optimizer, permitting its later updates, routing, and candidates to change in response.
4. Retain endpoint scores, trajectories, request usage, and failures. Evaluate contrasts on their declared common support and report the remaining planned observations.
5. For independent validation, fix each operation choice using the selection wave, then evaluate it on the disjoint validation wave. Resample states with all their nested outcomes together.

The construction checks are factuality, admissibility, isolation of the focal operation, and native continuation. The resulting inference concerns the stated states, alternatives, endpoints, and continuation regime. Local seeds record controllable randomness; provider outputs are retained separately.

A.2 IMPLEMENTED BOUNDARIES

GPTSwarm. Five warm-up updates produce the factual graph state. At the focal update, Native binds observed utilities to their original graph samples. CF is drawn uniformly from effective nonidentity bindings that change the utility tuple; states with empty effective support are excluded before downstream outcomes. This researcher-defined perturbation diagnoses the utility-binding update; it is not a proposed native search rule. Graph and optimizer parameters, observed utilities, and randomness required for restoration are retained. The intervention is followed by 15 native steps. The immediate displacement is $S_{L_2} = \|p^N_{\text{after}} - p^{CF}_{\text{after}}\|_2$; it uses both branches and is unsigned. The fresh study reuses each state's already frozen CF binding.

AFlow. The state is the restored workflow tree and its recorded guidance. Historical ordering analyses compare all committed legal parents; the two-horizon HumanEval comparison uses Native and the sole other committed parent. H_1 scores the generated child. H_2 executes another native selection and expansion, then scores that new child, whether or not it improves the incumbent. Selection can return to another existing node. HumanEval ordering uses the 14 of 21 states with tied guidance and three action-common repetitions; MBPP uses all 24 retained states.

MaAS. The frozen controller's first and second unique workflow draws define the alternatives for one focal query in a four-query batch; the other batch items remain matched. Each branch updates, runs one further native batch and update, then samples fresh probe workflows for four queries without updating again. The local signal is the difference in summed controller-layer log probabilities. The endpoint is $J = \text{score} - 3\,\text{cost}$ under the recorded nominal cost convention. Four warm-up batches precede the intervention.

EvoMAS. Each state is a unique task-conditioned selection history over a common clean pool of six seed configurations, with fresh empty memory. Native selection is called once and its ordered pair of parents is frozen. The alternative is selected before outcomes from legal ordered pairs with minimum overlap with Native, using a fixed seed. The selected tuple is forced on replay; parent adaptation and

three native evolution steps then execute. Model IDs are normalized to the same model stack. This evaluates an initial parent-selection boundary, rather than an intervention in independently evolved populations. The MMLU task answer is written to an audit sidecar and never fed back into search reward or selection.

A.3 Support, uncertainty, and horizon summaries

Common valid IDs define each completed-support contrast. Empty support gives an unavailable estimate. For scores in $[0, 1]$, a planned arm with sum s of observed scores, m missing scores, and R planned repetitions has finite mean bounds $[s/R, (s + m)/R]$. Propagating these bounds through the specified contrast retains the planned cohort. Such ranges describe missing outcomes, whereas bootstrap intervals describe sampling uncertainty on the retained states.

For a horizon comparison, $\widehat{\delta}_i = |\widehat{A}_{H_2,i}| - |\widehat{A}_{H_1,i}|$. The shared-ID sensitivity computes both endpoints on the same valid repetition IDs within each state. If a signed estimate is unbiased, Jensen's inequality still gives $\mathbb{E}|\widehat{A}_H| \geq |A_H|$; changes in finite absolute estimates can therefore include estimation variability. All completed-support and bounded-outcome summaries keep their respective support rules.

A.4 Conceptual implementation views

Figures 4 and 5 show the two-branch primitive and the original three implementation boundaries. Figure 1 combines those views and adds EvoMAS. Same-operation execution is defined in Section 3; independent selection validation is shown in Figure 2.

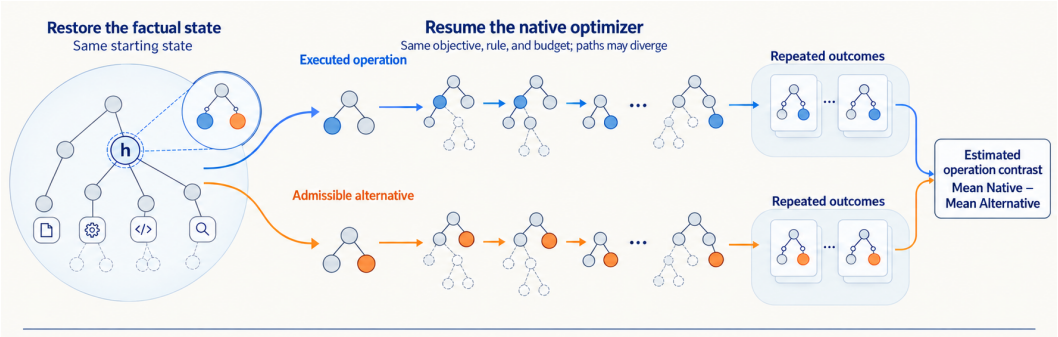


Figure 4: **The state-matched two-branch primitive.** The executed operation and its admissible alternative start from the same recorded state and then resume native search. Trees are conceptual, with no measured values encoded by position or color. Section 3 defines the additional same-operation control.

B GPTSwarm independent validation and complete controls

B.1 Frozen acquisition and evaluation design

The 20 states (IDs 0–19) were selected by availability of eligible frozen checkpoints before fresh outcomes. States 20–29 from the historical extension lacked these checkpoints and were not substituted. Each state contributes three arms and eight repetitions per arm, split into two waves of four. A three-branch pilot is excluded from inference. Every formal branch continues for 15 steps with four training executions per step and evaluates the final sampled policy on 153 frozen MMLU items. The training pool contains 285 items; exact-prompt overlap with the evaluation set is zero. The endpoint covers 49 subjects and the training pool 57. Inference conditions on the same fixed evaluation set in both waves.

The model is `openai/gpt-4o-mini-2024-07-18` through OpenRouter with the OpenAI provider, temperature 0.2 and top-p 1. Native node output caps are retained (50 for DirectAnswer and 1,000 for the default AdversarialAnswer). Isolated worker processes use branch-specific

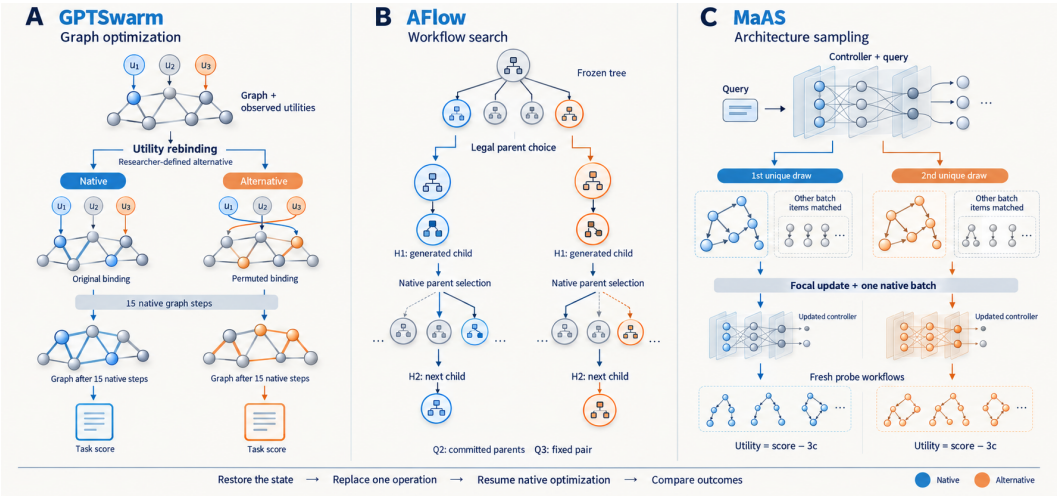


Figure 5: **Boundary-specific implementation views.** Graph utility binding, workflow-parent choice, and controller draws require different restored objects and endpoints. The AFlow panels distinguish immediate and subsequent generated children. MaAS evaluates fresh probes after the controller updates. These diagrams illustrate mechanisms; they contain no experimental trajectories or measured performance.

Python/NumPy and task/graph seeds. Native-A/B operation hashes match exactly; both use factual utility binding. CF uses the original frozen nonidentity binding. Historical post-update probabilities are checked within 10^{-6}; new branches match the current-runtime operation hash exactly. The three-attempt transport limit is retained, with every attempt charged or reserved. Failed scientific branches are never rerun for better outcomes.

All 480 branches complete, with 160/160 in each arm and all 20 states represented. Within each wave, state/arm/repetition slots are shuffled before acquisition; the two waves execute sequentially. The primary outcome is G_{validate} in Eq. 4; integer counts of correct answers determine signs and ties. State bootstrap uses 10,000 draws and seed 202609241346. Secondary signal and trajectory analyses retain their pre-wave-2 declarations and unadjusted intervals. Each resampled record keeps all repetitions and the first-wave choice together. These intervals summarize across-state variation in the observed cohort; they do not separately resample new questions or a new selection experiment.

B.2 Primary result and same-operation controls

Table 4 reports all aggregate outcome controls. The paired selection–validation gap is calculated before resampling, preserving state-level dependence. Same-operation selection chooses between Native-A and Native-B in wave 1 and evaluates the fixed choice in wave 2. The historical-choice analysis freezes the historical Native-A–CF sign and evaluates it on all eight fresh repetitions.

The signed same-operation offset is retained without seed reselection or outcome recentering. Its source is not isolated by this design. Absolute-contrast subtraction is a descriptive comparison; the experiment does not calibrate it as an unbiased estimate of expected-effect magnitude or test equivalence between the arms. The cross-operation selected advantage relative to the same-operation selection control is imprecise.

Across the two waves, 12 states preserve contrast sign, seven reverse, one has a selection tie, and none has a validation tie. Wave-rank correlation is 0.231 $[-0.231, 0.624]$. The score grid is $1/153 = 0.654$ pp per branch. Zero of 480 formal branches are perfect and zero score zero. The planned-cohort missingness bound collapses to the observed primary estimate because every planned branch completes; this adds no population confidence statement.

Table 4: **GPTSwarm outcomes and controls.** All values are accuracy percentage points; $N = 20$. Brackets are 95% state-bootstrap intervals. Only the independently validated Native-A/CF choice is the primary endpoint.

Quantity	Estimate	95% interval
Apparent selection advantage	2.173	$[1.634, 2.761]$
Independent validation advantage	0.654	$[-0.556, 1.822]$
Paired selection–validation gap	1.520	$[0.408, 2.696]$
Signed Native-B minus Native-A	-1.001	$[-1.965, -0.057]$
Signed Native-B minus CF	0.123	$[-1.083, 1.254]$
Absolute Native-B minus Native-A	1.973	$[1.397, 2.561]$
Absolute Native-B minus CF	1.993	$[1.254, 2.815]$
Paired difference of absolute contrasts	0.020	$[-0.895, 0.952]$
Same-operation selected advantage	0.899	$[-0.384, 2.190]$
Cross-operation minus same-operation selected advantage	-0.245	$[-2.002, 1.503]$
Historical choice, fresh eight-repetition advantage	-0.020	$[-1.021, 0.997]$
Mean within-state, within-arm standard deviation	3.585	$[3.295, 3.882]$

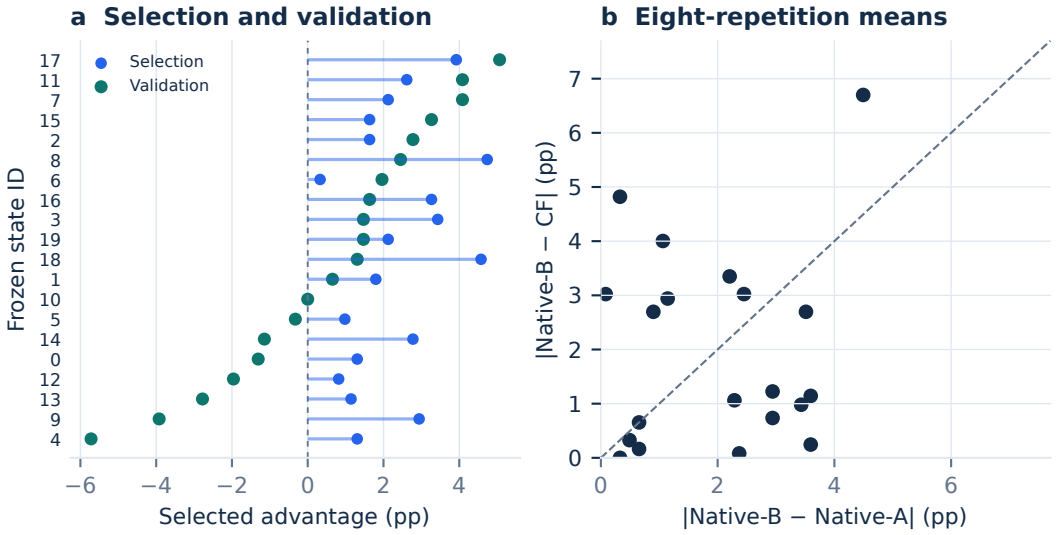


Figure 6: **All 20 states and the repeated-execution reference.** Left: apparent and independent selected advantages, ordered only for display by validation advantage. The first-wave tie contributes zero. Right: absolute differences of eight-repetition means using a common Native-B reference. Each point is one frozen state.

B.3 ALL LOCAL-SIGNAL COMPARISONS

All predictors in Table 5 use the fresh wave-2 Native-A–CF contrast as their target. Historical endpoint E requires a completed downstream evaluation, so it is distinct from immediate signals. The historical same-state Native-A–CF contrast has correlation -0.003 $[-0.511, 0.488]$ with wave 2; the historical independent-Native–CF comparison has 0.364 $[-0.165, 0.776]$. These secondary estimates retain all declared comparisons.

Table 5: **Local signals against independent outcomes.** $N = 20$; unadjusted 95% state-bootstrap intervals.

Signal	Spearman ρ	95% interval
Post-update edge-probability displacement	0.130	$[-0.348, 0.559]$
Pre-update factual batch mean	-0.112	$[-0.604, 0.394]$
Signed expected-edge-count difference	-0.172	$[-0.674, 0.355]$
Native gradient surrogate advantage	0.118	$[-0.311, 0.519]$
Historical Native terminal endpoint E	0.607	$[0.160, 0.864]$
Wave-1 immediate Native–CF score gap	0.081	$[-0.471, 0.568]$
Wave-1 immediate Native mean score	0.250	$[-0.193, 0.614]$

B.4 CONTINUATION-TRAJECTORY DISTANCE

The trajectory contains the four candidate-edge probabilities after native updates $t = 1, \ldots, 15$, excluding the immediate post-intervention $t = 0$. Distance is the mean over these steps of Euclidean probability-vector distance divided by $\sqrt{4}$, hence lies in $[0, 1]$. Equation 3 uses all cross-arm pairs and equally weights the two within-arm means. Results are computed for each wave ($R = 4$) and the pooled repetitions ($R = 8$).

Table 6: **GPTSwarm process measurements.** Units are normalized probability-trajectory distance, not accuracy.

Quantity	Estimate	95% interval
Pooled Native-A/CF separation S	0.006279	$[0.004270, 0.008255]$
Pooled Native-A/Native-B separation S	-0.000008	$[-0.000753, 0.000889]$
Paired excess separation	0.006288	$[0.004038, 0.008490]$
Wave-1 Native-A/CF separation	0.006726	$[0.003945, 0.009715]$
Wave-2 Native-A/CF separation	0.005155	$[0.002156, 0.008256]$

The wave-1 separation–wave-2 value association is 0.000 $[-0.453, 0.471]$. The rank association of separation across waves is -0.029 $[-0.495, 0.450]$. These describe individual-state ranking, separately from positive average process separation.

B.5 COMPLETION, RECOVERY, AND RESOURCES

Three session recoveries affect 36 distinct interrupted branches. Saved responses are replayed exactly to reconstruct the completed prefix before continuing; already completed branches are not rerun. Operation, input, and reply validations preserve the frozen acquisition and analysis definitions. A later operational annotation corrects the recovery count from the first 12 marked branches to the complete 36; it changes no scientific estimate.

Table 7: **Formal GPTSwarm resource receipts.** Unknown-charge attempts retain full reservations. Request durations are summed across concurrent requests and are not elapsed wall time.

Arm	Attempts	Input tokens	Output tokens	Receipt US$	Upper US$
Native-A	47,895	10,539,368	284,960	1.7519	1.9284
Native-B	48,557	10,693,463	285,717	1.7754	2.2031
CF	47,494	10,426,532	261,665	1.7210	2.0524

Formal receipts cover 143,898 successful requests; 48 other formal attempts have unknown charges. The excluded pilot has 848 receipted requests. Total attempts are 144,794, confirmed costs are US$5.2792632, and the accounting upper bound is US$6.2148432. Formal input/output totals are 31,659,363 and 832,342 tokens. Recorded request durations are 38,122, 39,017, and 37,758 seconds for Native-A, Native-B, and CF; missing timing counts are 9, 14, and 13. Approximately 13.97 hours elapse from start to completion, including session interruptions. Per-branch tokens, calls, and charges reconcile to the ledger.

C EvoMAS: process, reward, and task endpoints

C.1 Separate datasets and evidence levels

BBEH-mini and MMLU are separate cohorts. BBEH plans 20 states and excludes protocol-invalid state 012 under its frozen rule, leaving 19. Native and CF each have three repetitions. Its retained evidence comprises a frozen state table, analysis code, and provenance records. MMLU plans 30 task-conditioned states and retains 210 raw records (one factual run and six replay slots per state). Twenty-four states have complete three-per-arm trajectories and reward outcomes; 20 have complete parsed task endpoints. No WorkBench, calendar, or email run is included as empirical evidence.

MMLU uses the historical EvoMAS reasoning pool with model-ID normalization, $K = 2$ parents, $T = 3$ steps, one evaluation task per state, and fresh memory. The meta-model temperature is 0.7 and judge temperature 0. Native search reward is

$$J = J_{\text{judge}} - 10^{-6}(\text{tokens} + 1000\,\text{timeSeconds}).$$

MMLU exact-match correctness is a deterministic audit sidecar, invisible to search. It was the frozen primary task endpoint; native reward was secondary. The recovered raw records reproduce the historical state-level metrics. Offspring-only separation and its missingness/resource analyses are additional offline sensitivities, computed without new model calls.

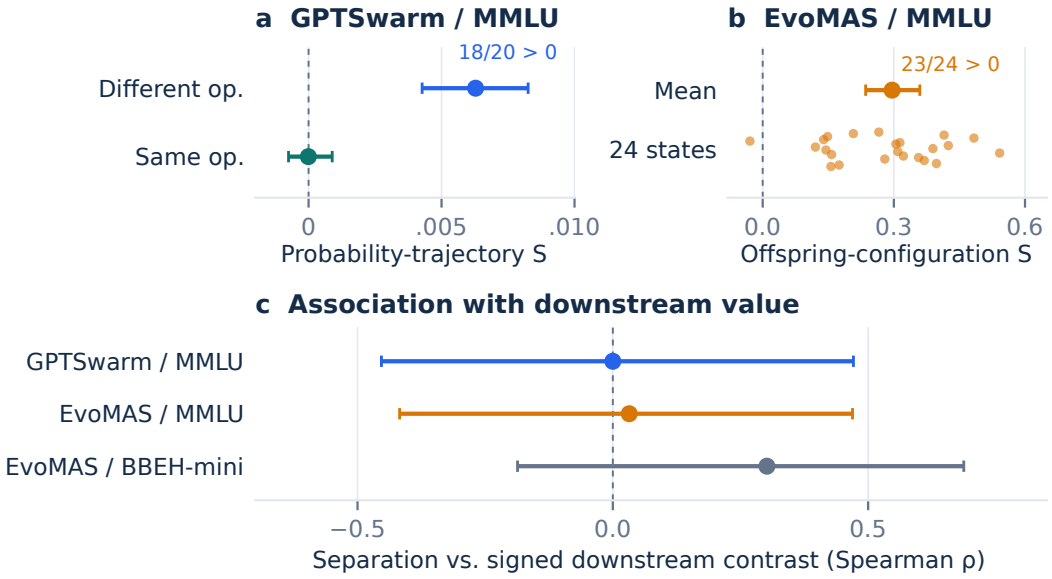


Figure 7: **Process separation and downstream association.** Panels (a) and (b) use framework-specific distance scales: GPTSwarm compares post-intervention probability trajectories, and EvoMAS/MMLU compares subsequent offspring configurations. Panel (b) shows individual states; points with bars show means and 95% state-bootstrap intervals. Panel (c) relates unsigned separation to the Native-relative signed contrast. GPTSwarm uses wave-1 separation against wave-2 accuracy; EvoMAS uses secondary native reward. BBEH uses its frozen full-trajectory measure. Broad association intervals accompany positive mean process separation.

C.2 TRAJECTORY DEFINITION AND DISTANCE AUDIT

Canonicalization removes runtime metadata, normalizes model identifiers, sorts mappings, and preserves semantic fields and prompt content. Configuration distance is the fraction of unequal flattened fields over the union of field paths. It therefore includes prompt and configuration changes, rather than only topology.

The historical full-trajectory distance averages three components: optimally matched adapted-parent configuration distance, average offspring configuration distance over aligned evolution steps, and action-sequence mismatch. The MMLU offspring-only sensitivity retains only the second component, excluding adapted parents and action labels. With three repetitions in each arm, B averages nine cross-arm distances and W averages six within-arm distances (three per arm). Both use the same distance function before taking $S = B - W$.

Table 8: **EvoMAS process and value evidence.** BBEH uses its frozen full-trajectory distance; MMLU also supports offspring-only reconstruction. Reward units and distance scales are kept distinct.

Quantity	MMLU ($N = 24$)	BBEH-mini ($N = 19$)
Positive full-trajectory S	24/24	19/19
Mean full-trajectory S	0.22365	0.22714
Positive offspring-only S	23/24	Unavailable
Mean offspring-only S	0.29600	Unavailable
95% interval	$[0.23549, 0.35936]$	—
Mean Native–CF reward	-0.11434	-3.00772
95% interval	$[-3.54666, 3.06025]$	$[-10.92191, 3.82957]$
Separation–reward ρ	0.03217 (offspring)	0.30175 (full)
95% interval	$[-0.41755, 0.46941]$	$[-0.18657, 0.68728]$

The historical full-trajectory MMLU separation has mean interval $[0.18954, 0.25863]$ and reward correlation -0.265 $[-0.636, 0.183]$. Both full and offspring-only results are retained. MMLU reanalysis uses 10,000 state resamples with seed 20260925; BBEH preserves the original frozen resampling rules and reproduces all six reported intervals.

C.3 TASK RESOLUTION AND REWARD DECOMPOSITION

All 20 complete MMLU task contrasts equal zero. Fifteen states have both arms perfect, four have both arms at zero, and one has both arms at $2/3$. A correlation with this constant contrast is undefined. Thirty planned states, a single task per state, and incomplete answer parsing provide limited information about task-level operation value; the planned task-contrast range allowing missing answers in $[0, 1]$ is $[-0.05556, 0.07778]$.

The reward decomposition is observable on all 24 complete reward states. The mean judge-score contrast is -0.09722 $[-3.54167, 3.12500]$; the resource-penalty contribution is -0.01712 $[-0.03976, 0.00794]$. Their sum is the reward contrast -0.11434. Fifteen states have tied mean judge scores, so all their nonzero reward differences arise from resource penalties; the largest absolute difference within this group is 0.04093. These values are native reward units, not accuracy percentage points.

BBEH's separately recorded E evaluates the factual final configuration three times: 13 states have $E = 0$, four have $E = 1/3$, and two have $E = 1$. These factual measurements do not provide a Native–CF task-accuracy contrast. The full-trajectory/reward analysis in Table 8 is supported by the state table; additional branch-level or offspring-only BBEH results are unavailable.

C.4 COMPLETION AND INFLUENCE SENSITIVITIES

Allowing six incomplete MMLU states arbitrary $S \in [-1, 1]$ gives planned-30 mean offspring-separation bounds $[0.03680, 0.43680]$. These are finite missingness bounds. Deleting one complete state at a time gives mean offspring separation in $[0.27788, 0.31014]$.

Table 9: **MMLU branch completion and answer parsing.** Parse failures remain missing for task correctness. Runtime completion and task parsing are distinct counts.

Arm	Planned	Runtime complete	Scientific failure	Parseable	Unparseable
Native	90	86	4	84	2
CF	90	87	3	84	3

For BBEH, the frozen sensitivity excludes repaired states 005, 007, 010, 011, and 019. The remaining 14 all have positive full-trajectory separation; mean reward difference is 0.30676 $[-8.28257, 6.81317]$. Across all 19 single-state deletions, mean reward contrast ranges from -4.37334 to -0.63294 and the separation–reward correlation from 0.22807 to 0.43653. All original states remain in the main 19-state result. The original per-state classification gives 17 unresolved, one beneficial-resolved, and one nonbeneficial-resolved state; the beneficial-resolved contrast is only 0.0360 reward units. Resolution of a sign and practical effect size are reported separately.

D HISTORICAL LOCAL-SIGNAL EVIDENCE

D.1 GPTSWARM TASK AND BACKBONE COMPARISONS

The historical utility-binding intervention has nonzero immediate displacement in 28/30 GPT-4o-mini MMLU states and 27/30 GPQA-Diamond states. Native-relative orientation is fixed throughout: exchanging arm labels leaves the unsigned displacement unchanged but reverses the signed value contrast. Figure 8 preserves all retained states.

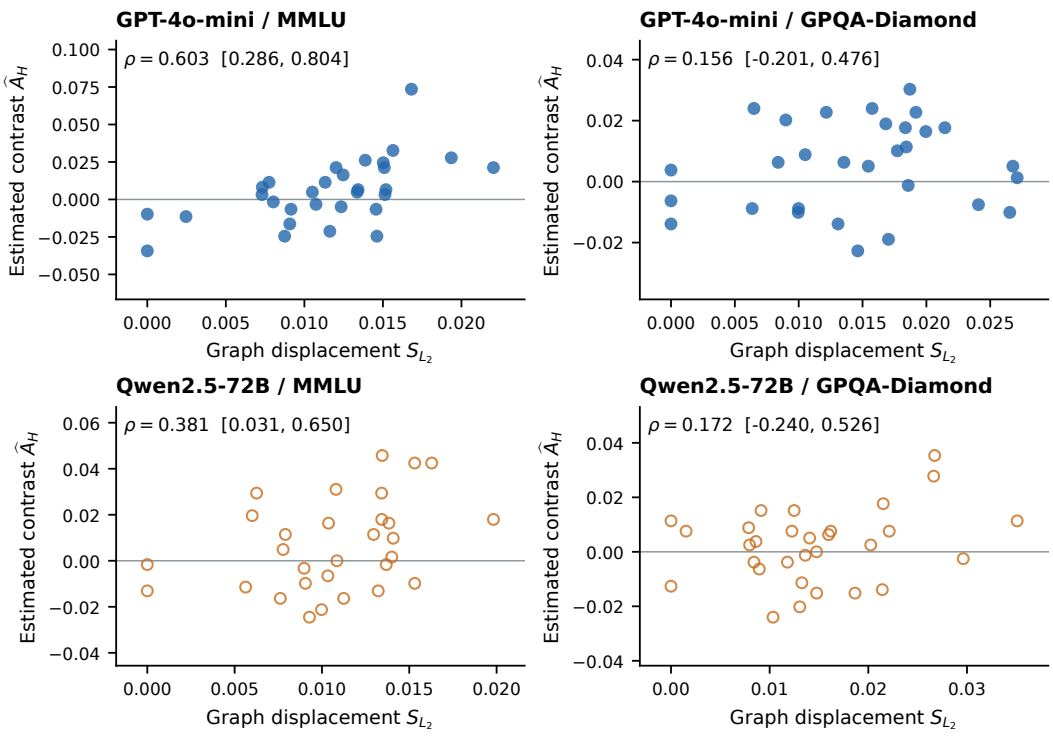


Figure 8: **Historical local measurements and downstream contrasts.** Each panel contains 30 states. Horizontal position is immediate edge-probability displacement; vertical position is the signed contrast after 15 native updates. Qwen is secondary; its MMLU contrasts were reconstructed from retained endpoint records. The independent fresh validation is reported separately in Appendix B.

The direct MMLU–GPQA correlation difference is 0.44747 $[-0.00908, 0.85898]$ for GPT-4o-mini and 0.21 $[-0.28, 0.69]$ for Qwen. Historical MMLU combines 20 initial states with an extension

Table 10: **Historical associations and acquisition sensitivity.** Original-cohort intervals retain their original computation. The last three rows use separately regenerated acquisition intervals.

Cohort	N	ρ	95% interval
GPT-4o-mini / MMLU, historical	30	0.60319	$[0.28556, 0.80358]$
GPT-4o-mini / GPQA-Diamond	30	0.15571	$[-0.20112, 0.47603]$
Qwen / MMLU, secondary	30	0.38	$[0.03, 0.65]$
Qwen / GPQA-Diamond, secondary	30	0.17	$[-0.24, 0.53]$
MMLU initial states 0–19	20	0.723645	$[0.370368, 0.919333]$
MMLU extension states 20–29	10	0.352585	$[-0.468354, 0.858607]$
MMLU combined, acquisition audit	30	0.603186	$[0.283948, 0.805878]$

frozen after interim analysis. Acquisition-audit intervals use 10,000 draws with seeds 20260920–20260922 for initial, extension, and combined cohorts. Deleting one MMLU state gives correlations 0.56202–0.67078 under GPT-4o-mini and 0.32146–0.45428 under Qwen. These overlapping subsets diagnose influence. Qwen supplies 30 eligible MMLU states from 32 attempts and 30 GPQA states from 40; eligibility is established before downstream counterfactual outcomes.

D.2 Other logged signals

Historical exploratory comparisons retain the full candidate list in the source package. In the ten-state MMLU extension, factual utility, signed branch displacement, Native update norm, and signed Native update have correlations $-0.033, -0.195, -0.097, -0.170$, respectively. In GPQA, factual utility, signed branch displacement, Native focal loss, and CF-minus-Native loss have $-0.161, 0.270, -0.312, 0.442$. All direct differences between displacement's correlation and the comparator correlation have intervals containing zero. GPQA's loss-margin interval is $[0.152, 0.640]$, an unadjusted exploratory association. The new 20-state independent signal comparisons, including immediate score evaluations, are all reported in Table 5.

D.3 MaAS objective decomposition

The 29 MaAS runs contribute three repetitions and four probe workflows per arm (696 endpoint records). The recorded local margin is the difference in summed controller-layer log probabilities for the focal draws. Mean utility contrast is 0.00560 $[-0.01188, 0.02583]$ under the historical interval, with margin association 0.29163 $[-0.11026, 0.65295]$. Each recorded probe satisfies $J = \text{score} - 3c$. Table 11 separates the components; its regenerated intervals are supplementary.

Table 11: **MaAS score and cost decomposition.** Differences are Native minus CF. Recorded cost is nominal USD-priced token usage, separate from study acquisition cost.

Quantity	Native	CF	Difference	Supplementary 95% CI
Score	0.890805	0.885057	0.005747	$[-0.011494, 0.025862]$
Cost	0.002285	0.002235	0.000051	$[-0.000025, 0.000128]$
Utility	0.883949	0.878354	0.005595	$[-0.011900, 0.025834]$

Twenty runs have zero mean score contrast but nonzero utility contrast due to cost. The mean score difference 0.00574713 plus the cost-penalty contribution -0.00015186 gives utility difference 0.00559527. At coefficients $\lambda = 0, 1, 3, 5$, mean $\text{score} - \lambda c$ contrasts are 0.005747, 0.005697, 0.005595, and 0.005494; margin correlations are 0.017, 0.292, 0.292, and 0.292. Every corresponding mean and association interval spans zero. Full coefficient tables remain in the accompanying archive. This reweights recorded endpoints and does not retrain the optimizer.

No probe index overlaps optimization query indices within any run: 348 paired probe slots are nested in 29 runs. All derive from the finite 164-task HumanEval pool. This within-run separation is distinct from a separately sampled generalization test set.

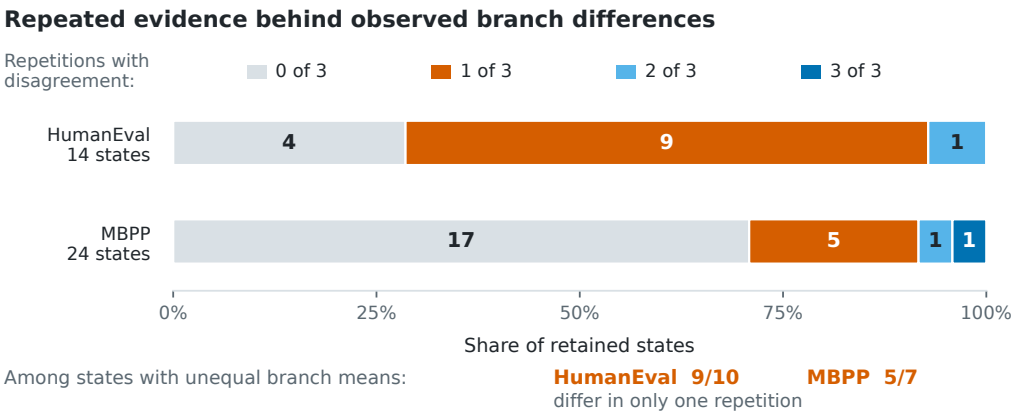


Figure 9: **Repeated support behind observed branch differences.** Bars partition retained tied-guidance states by the number of three common repetitions with cross-action disagreement. Segment labels are state counts; widths are within-task proportions. Among states with unequal branch means, 9/10 on HumanEval and 5/7 on MBPP differ in only one repetition. These are descriptive counts, not tests of expected action ordering.

E AFlow repetition, continuation, and missingness

E.1 Historical repetition and score resolution

Historical AFlow uses at most three outer repetitions and one scoring pass per generated child on ten fixed items. The tied-guidance ordering cohorts contain 14 HumanEval states and 24 MBPP states. Common valid IDs are intersected across all committed actions; a cross-action range above 10^{-12} marks disagreement. Of the states with different full-sample means, 9/10 HumanEval and 5/7 MBPP states differ in just one repetition.

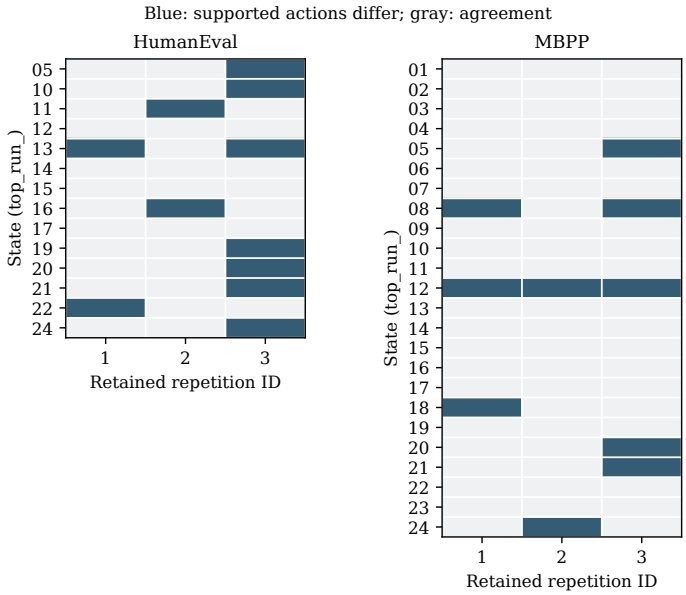


Figure 10: **Every retained ordering state.** A colored cell indicates cross-action disagreement at that repetition ID. Rows are states; repetitions remain nested observations. This view preserves the raw support behind each observed branch-mean difference.

HumanEval contributes 84 scored workflows: one at 0, 14 at 0.7, and 69 at 0.8. MBPP contributes 144: six at 0, four at 0.6, and 134 at 0.7. Neither cohort reaches 1.0. At least one action varies across repetitions in 11/14 HumanEval states and 7/24 MBPP states. The 0.1 score grid limits discrimination even below the full-score ceiling.

The leave-one-repetition-out diagnostic fixes an ordering on two repetitions and evaluates it on the third. For two actions it is $R^{-1}\sum_r \mathrm{sign}_\epsilon(\overline{D}_{-r})D_r$, with $\epsilon = 10^{-12}$ and a zero contribution for a training tie. If only one repetition differs, each fold has either a training tie or a zero evaluation contrast. Thus a zero diagnostic follows mechanically in that case. Positive state-mean diagnostic counts are 1/14 and 2/24, with overall means 0.00476 and 0.03750. This reuses observations; the fresh experiments supply independent repetitions.

The GSM8K allocation comparison uses pre-outcome Native and uniform weights over the same parents. Nineteen of 20 states have nonuniform Native weights; 17 have common valid outcomes. Mean weighted contrast is 0.02471 $[0.00000, 0.06588]$. The three missing states remain in the planned accounting. This evaluates fixed mixtures, separately from selecting the empirically best parent.

E.2 Historical horizon comparison and influence

The 21 HumanEval states give mean absolute contrasts 0.03889 at H_1 and 0.01429 at H_2. Their paired change is -0.02460 $[-0.05397, -0.00476]$; using common IDs across both horizons gives -0.02302 $[-0.05238, -0.00317]$. The primary change decreases in nine states, is zero in ten, and increases in two. Shared-ID counts are eight, eleven, and two. Both medians equal zero.

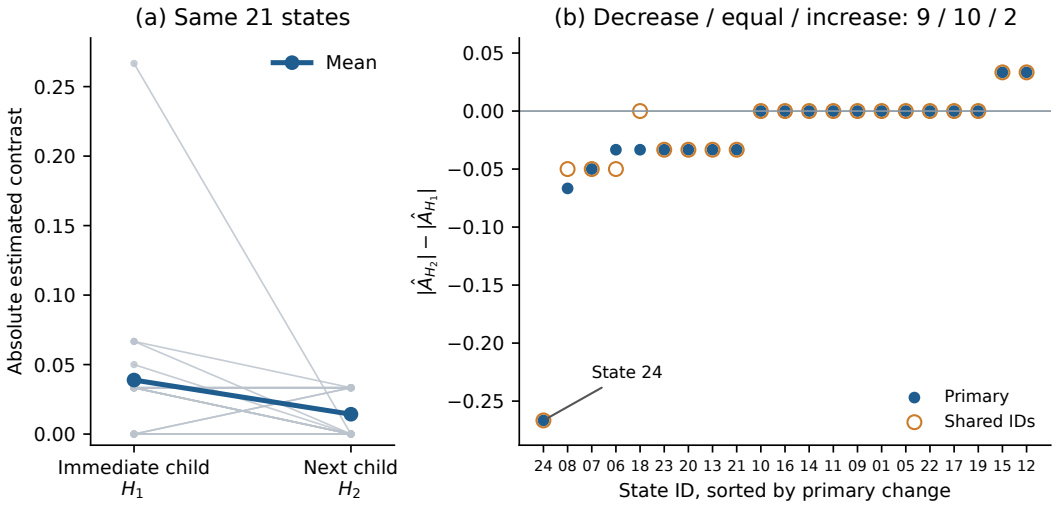


Figure 11: **Historical state-level continuation sensitivity.** The left panel pairs immediate and next-child magnitudes; the right retains all state changes under primary and shared-ID support. Coincident states can overlap. The endpoint is the generated child's score at each horizon.

Single-state deletion gives primary mean changes in $[-0.02750, -0.01250]$ and shared-ID changes in $[-0.02583, -0.01083]$. The largest individual decrease, -0.26667, is state 24. Removing it gives -0.01250 with supplementary interval $[-0.02333, -0.00167]$, and shared-ID -0.01083 $[-0.02167, 0.00000]$ (50,000 draws; seeds 2026092401 and 2026092402). All states remain in the main comparison. Of 117 completed next expansions, 83 select another node and 34 select the immediate child; 14 next-child scores fall below the recorded incumbent.

The 126 planned trajectories supply 121 observed H_1 and 117 observed H_2 scores, with 59 and 55 matched pairs. Bounding every missing endpoint in $[0, 1]$ gives planned mean horizon-change bounds $[-0.07302, 0.08095]$. Defining zero operational utility for structural failures while bounding provider-missing outcomes gives $[0.00159, 0.05714]$. This second scenario changes the endpoint definition; neither range is a confidence interval. Structural failures include four missing-operator

Table 12: **Historical arm-specific completion.** Failure rows partition incomplete trajectories; an H_1 score can survive later failure.

Accounting	Native	CF
Planned trajectories	63	63
Complete through H_2	57	60
Provider failure	1	2
Structural failure at H_2	3	1
Structural immediate failure	2	0
Observed H_1 score (overlapping)	60	61

AttributeErrors and two float/string TypeErrors, whose individual causes are not isolated by these records.

E.3 Two new three-arm waves

The same 21 states are revisited with Native-A and Native-B independently expanding the factual parent, and CF expanding the other parent. Each trajectory scores H_1 and performs one further native expansion for H_2. The first wave has six repetitions; a second six-repetition wave is frozen after inspecting the first wave, without effect-based state or slot selection. Wave 2 is analyzed separately; twelve-repetition pooling is secondary. The same current scientific worker, scorer, provider, and tasks are retained between the two new waves.

Calls use the OpenAI model route, temperature 0, top-p 1, and a 4,096-token cap. Two excluded engineering pilots precede formal collection. An import-cache error is repaired after the first; the revised gate requires completion in each arm and reviewed failure classification. Native generation retries and workflow repairs remain charged operations. The current scorer validates all 33 canonical solutions, while the historical scorer's exact hash remains unavailable. Historical and new outcomes are not pooled.

Table 13: **Three-arm completion by new wave.** Each arm plans 126 trajectories per wave. Structural and operational counts concern terminal trajectories; finalized endpoints can survive later failures.

Wave	Arm	Planned	Final H_1	Final H_2	Structural	Operational
1	Native-A	126	119	104	11	11
1	Native-B	126	121	111	8	7
1	CF	126	125	114	6	6
2	Native-A	126	123	111	9	6
2	Native-B	126	124	110	7	9
2	CF	126	122	115	4	7

Nineteen first-wave trajectories are interrupted, retaining 11 verified H_1 checkpoints. They are not replaced. Concurrency changes and reserved unknown charges are recorded. The second wave's terminal failures comprise 22 ProviderAbort and 20 AttributeError records. Five first-wave TypeError-terminated H_1 expansions retain all scored items; an item-complete sensitivity preserves them separately from the primary native-finalization rule. At each horizon, common complete triples determine both contrasts, and shared-horizon analyses additionally intersect IDs across horizons. Intervals use 20,000 state draws with seed 20260923.

First-new-wave choices at H_1 have 5 preserved, 3 reversed, 5 new ties, and 8 selection ties in the added wave; at H_2 these counts are 3, 4, 7, and 7. Selection ties contribute zero advantage. Complete triples per state range 4–6 and 1–6 at H_1/H_2 in wave 1, and 4–6 and 2–6 in wave 2. Shared-support mean absolute Native–CF changes are 0.1263 $[0.0451, 0.2228]$ and -0.0180 $[-0.0530, 0.0140]$; corresponding Native–Native changes are 0.1020 $[0.0299, 0.1871]$ and -0.0387 $[-0.0873, -0.0010]$. Pooled-twelve changes are 0.0404 $[0.0076, 0.0775]$ and 0.0331 $[-0.0151, 0.0825]$. Full signed/absolute interval tables remain in the source archive. Planned Native–CF change ranges are $[0.0008, 0.2500]$ for wave 1, $[-0.0627, 0.1286]$ for wave 2, and $[-0.0484, 0.1726]$ pooled.

Table 14: **New-wave absolute contrasts and independent choices.** Accuracy-proportion units; intervals accompany the key independent-choice endpoint. Complete-triple support is horizon-specific in the first four rows.

Quantity	Wave 1 H_1	Wave 1 H_2	Wave 2 H_1	Wave 2 H_2
Mean $\lvert N_A - CF \rvert$	.0140	.1440	.0387	.0415
Mean $\lvert N_A - N_B \rvert$	.0220	.1180	.0450	.0160
Mean signed $N_A - CF$	.0045	-.0311	-.0267	.0372
Mean signed $N_A - N_B$	.0058	-.0131	-.0009	.0016

Choice fixed on	Evaluated at	Added-wave advantage [95% interval]
Historical ordering	H_1	.0006 $[-.0198, .0187]$
Historical ordering	H_2	.0009 $[-.0158, .0179]$
First new wave	H_1	-.0116 $[-.0361, .0063]$
First new wave	H_2	-.0107 $[-.0280, .0032]$

E.4 Additional-item evaluations

Original parents and first-repetition generated children are held fixed and reevaluated on 30 tasks: the original ten plus 20 disjoint tasks from the same 33-item validation pool. Parent scores measure fixed candidate performance, while generated-child scores measure fixed outputs of earlier search. These checks add no native optimization step. Child selection is by state/arm/repetition ID, not observed performance. The additional-item studies are secondary.

All 126 planned parent evaluations have terminal records: 114 complete, 11 operational failure, one hard stop. All 63 H_1 child slots terminate: 62 native-complete and one structural failure with all 30 item scores retained. Offline rescoring confirms those retained items. The combined-30 Native-A–CF parent difference is 0.0024 $[-0.0032, 0.0085]$ over 21 states; the first-child difference is -0.0016 $[-0.0111, 0.0079]$. Parent matched-repetition counts are one in two states, two in eight, and three in eleven.

Of 63 selected H_2 slots, 62 retain generated source and 42 have all 30 scored items. Terminal classes are 42 complete, 11 operational failure, one unavailable source child, four structural failure, and five hard stops. Only eight states have complete combined-30 three-arm support. Their signed Native-A–CF contrast is -0.1042 $[-0.3208, 0.0167]$, and the fixed-child shared $H_2 - H_1$ absolute change is 0.1125 $[-0.0042, 0.3250]$. The paired additional-20 minus original-10 signed contrast is -0.0250 $[-0.0813, 0.0250]$. Planned combined-30 signed bounds are $[-0.3524, 0.1841]$. Separate collection waves and incomplete support qualify these fixed-child comparisons.

E.5 Resource record

Including excluded pilots, AFlow's new studies have US$13.2610 in recorded charges and an accounting upper amount of US$26.6887 when all unresolved reservations remain charged. The initial and added three-arm waves use 37,957 and 38,870 requests, with US$5.0174 and US$4.9883 in receipts. Item-evaluation phases, actual tokens, summed request times, failures, and acquisition-segment budgets remain in the archived ledgers and machine-readable supplement. Historical nominal estimates and current receipts are different accounting scopes.

F Reproducibility and provenance

F.1 Available records and audit scope

The source package preserves frozen summary tables, branch/state results, figure-generating code, and source hashes. The new GPTSwarm archive restores all 480 branches from its base checkpoint and final delta, with raw replies, operation hashes, evaluation records, and the cumulative SQLite ledger. Saved-prefix validation checks both scientific inputs and replies before continuing interrupted executions. Bootstrap intervals resample the observed state records while keeping the evaluation items fixed.

EvoMAS/MMLU has 210 retained raw rows, a frozen protocol, historical distance functions, and reproducible state metrics. BBEH has a frozen state table and analysis code; six historical intervals are reproduced, but complete raw BBEH branches are unavailable. The two data sources are indexed separately. Offspring-only MMLU distance and reward decomposition are explicitly offline secondary analyses.

Historical construction documentation covers 154 canonical units across six cohorts (AFlow: 21 HumanEval, 24 MBPP, 20 GSM8K; GPTSwarm: 30 MMLU and 30 GPQA; MaAS: 29 HumanEval). Three GSM8K units have no computable estimate. These are documentation counts, distinct from successful reexecution. The qualified AFlow offline restoration check restores all 21 states and verifies 60/63 Native slots; three are untestable. It substitutes recorded generation/evaluator outputs and stops before a new H_2 model call. The exact historical scorer remains unrecovered.

Historical GPTSwarm source specifies a fresh graph and optimizer per state; its manifest does not fully attest every historical execution version. The new checkpoint study adds explicit operation hashes and retained replies. API seed schedules document local control; no analysis asserts bitwise provider reproducibility. Older and current acquisition regimes remain separate in the tables and figures.

F.2 Resource scopes

New GPTSwarm resource receipts are in Table 7; AFlow current receipts are summarized in Appendix E. For comparison, historical GPTSwarm MMLU initial/extension records have 72,427/36,844 calls, 15,926,553/8,121,460 input tokens, and 420,542/218,884 output tokens. Their US$2.6413/1.3495 costs use the logged nominal rates. Historical GPQA's retained 30-state records contain 29,680,383 input and 848,231 output tokens, giving a nominal US$4.9610; a comparable call count is unavailable. These nominal scopes are not reconciled billing receipts or matched incremental evaluation costs.