
Same Algorithms, Different Scientist: Design Space Bias in AI Discovery

Odysseus Kaloidas
UC Berkeley

Abstract

1 AI scientists often choose from human-written menus of methods. We ask whether
2 rewriting such a menu changes measured scientific competence when every exe-
3 cutable option remains available. We group policies that query identical variant
4 sets on fixed probe states into an *executable behavioural quotient*. These groups let
5 us test selection stability and measure discovery credit relative to uniform selection
6 over distinct behaviours. Across 12 combinatorial protein-fitness landscapes, 40–41
7 executable policies collapse to 21–22 distinct campaigns; 19 written options imple-
8 ment one algorithm. Nevertheless, equivalent rewritings move model choices by
9 a median total-variation distance of 0.417, against a 0.144 same-prompt replicate
10 floor, and reverse 12 of 85 primary-family credit verdicts. A separately sealed
11 knapsack grammar reproduces order sensitivity across 24 permutations and 6 prob-
12 lem families. Controls distinguish positional and textual preferences: probability
13 falls with menu position even after controlling for utility, while the shared textual
14 ordering is negatively associated with utility. Duplicating one default lowers selec-
15 tion AUC from 0.854 to 0.753 for one chooser and raises it from 0.199 to 0.710
16 for another. Presenting one entry per behaviour stabilises measurement without a
17 detected pooled discovery gain. Evaluations should report behaviour-level credit
18 and utility separately: removing a representation bias can also remove a useful
19 prior.

20 1 Introduction

21 Language models propose optimisers [7, 6, 9], experiments [1], and papers [14]. A human first
22 specifies what they may propose: a program skeleton, grammar or menu. Crediting the model
23 for selecting from these options requires separating its contribution from the curator’s choice of
24 representation.

25 Consider choosing a discovery policy on the DHFR protein landscape, with utility measured as mean
26 best-found fitness. The standard policy and 18 differently written options query the same variant sets
27 on our probe states, so we count them as one behaviour, not 19 opportunities. We define discovery
28 credit as the chooser’s expected utility minus that of sampling uniformly over distinct behaviours.
29 In one observed example, 23 of 24 Claude Sonnet 4 replies select a Gaussian-process policy and
30 one selects diverse batch selection. Their mean utility is 4.021 versus the uniform reference’s 3.895,
31 giving credit +0.126. Reordering the same options makes all 24 replies select expected improvement,
32 with utility 3.247 and credit -0.649. The evaluation changes from above to below its reference
33 although no available policy changed.

34 Our test changes menu order, writes equivalent options as separate entries (aliases), or adds settings
35 that have no effect on execution. We call changes in choice probabilities under this fixed executable
36 set *design-space representation bias*. Repeated identical prompts provide the noise reference.

Table 1: Evidence populations and their roles. A choice contrast compares two distributions; a logit cell is one model, landscape and arm. BQN presents one entry per behaviour class. Repeated utility contexts are not independent choice prompts. Failed predictions appear in Table 10.

result / protocol	population and eligibility	endpoint and evidence boundary
Execution audit	41 declared policies including the default; 40–41 execute on 12 protein landscapes	21–22 classes; independent grammars: 28–29/34 and 27–28/37; AutoML: 12–18/29
Sampled choice	4 hosted models; 8,088 calls, 31 invalid replies; 41 canonical model–landscape cells; at least 8 valid replies per arm for credit	85 primary contrasts: median TV 0.417, 44 significant; 12/85 credit flips. Extended primary: 38/193; secondary: 18/62
Logit round	7 models $\times$ 12 landscapes $\times$ 14 arms = 1,176 cells; 27 excluded by the mass floor, all arrangement arms; none missing	3 distinct prompt groups; local logits validated, hosted logits sensitivity-only; 84 canonical/raw pairs, 645 eligible ensemble cells
Position and repair	Position: 84 cells on two identifier-coded arms; BQN: 84 paired cells	Position controls utility; BQN AUC +0.010 [−0.046, +0.086]: no detected pooled gain, not equivalence
Duplication	2 preselected models; 12 landscapes; 5 copy counts; held-fixed class utilities	Opposite slope signs; broader sign prediction and AutoML transport fail

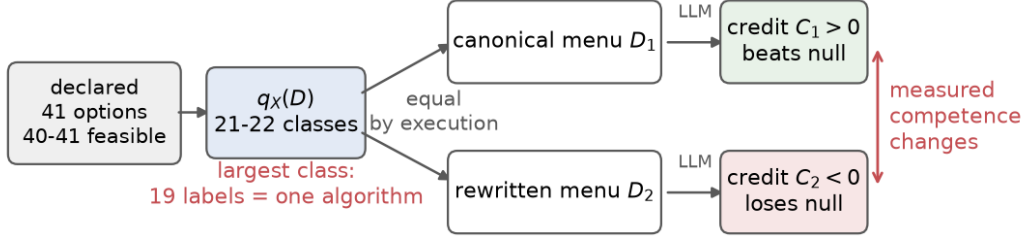


Figure 1: Testing representation while holding behaviour fixed. Both menus offer the same behaviours, verified by executing every option on the probe states. Model choices are grouped by behaviour before computing credit against uniform selection over those groups. A byte-identical prompt replicate measures variation that occurs without changing the menu.

Figure 1 summarises the comparison. Prior work establishes tool-selection bias [24]; execution checks here distinguish stable credit from useful selection.

Table 1 distinguishes the execution audit, model-choice experiments and repair tests.

2 Experimental design and measurement

Discovery task and behavioural equivalence. A *discovery policy* runs a closed-loop campaign on a combinatorial protein fitness landscape: 24 initial observations, then four adaptive rounds selecting 24 unmeasured variants each—96 adaptive assays and 120 observed variants in total—scored by the best fitness revealed. A policy specifies a representation, a predictive surrogate, a variant generator, an acquisition rule, batch selection, updating and stopping. These are its changeable *coordinates*. Starting from a standard configuration, we change one coordinate at a time to obtain 41 declared policies, tested in every arm on 12 completely measured landscapes. One recombination policy is infeasible on 9 landscapes, leaving 40–41 executable policies per landscape. For frozen probe states X the signature $\sigma_X(\pi)$ is the sequence of variant *sets* the policy queries; $\pi \sim_X \pi'$ when $\sigma_X(\pi) = \sigma_X(\pi')$ exactly, inducing $q_X : D \rightarrow D/\sim_X$. The quotient groups policies with identical query sequences. Grouping uses *selections*, not fitness: equal outcomes alone would not establish equal behaviour (488 fidelity checks; Appendix I).

With task information S fixed, representation invariance requires model M to assign the same probabilities to behaviours under equivalent menus D_1, D_2 :

$$P_M(q_X(\pi) \mid S, D_1) = P_M(q_X(\pi) \mid S, D_2) \quad \text{whenever} \quad q_X(D_1) = q_X(D_2). \quad (1)$$

Menu rewrites and credit. A *written space* is a menu naming executable policies, described with a common phrasing function. The primary rewrites change order, expand aliases or add inert coordinate changes (*dead inflation*). Secondary rewrites label the default, replace descriptions with neutral codes, or lengthen them. We separate these because they can change useful textual information even when execution is unchanged (Appendix H). Execution verifies that every constructed space has the same quotient as the canonical one. A model is shown one written space and asked for one identifier, with the first identifier shown as an answer-format example. Positional effects therefore include this cue. Everything but the menu text is fixed across arms; a *canonical replicate* arm issues the byte-identical prompt as a separate batch, so no distance is judged against zero.

Before any model call, we execute each class B on 20 evaluation seeds (100–119), disjoint from the probe seeds; its utility $u(B)$ is mean best-found fitness. We add probabilities of entries in the same class, then compare expected utility with uniform selection over classes rather than written entries: $C(M, D) = \mathbb{E}_{B \sim P_M(\cdot \mid D)}[u(B)] - \mathbb{E}_{B \sim \text{Unif}(q_X(D))}[u(B)]$. Positive *discovery credit* means outperforming this *behaviour-uniform* reference, as in the DHFR example (Section 1).

Repeated policy selection. We also ask what utility a chooser reaches when allowed several independent policy draws, retaining the best. Selection AUC is the area under this best-utility curve at $n \in \{1, 2, 3, 4, 6, 8, 12, 16, 24\}$ draws from the measured class distribution. For utilities sorted increasingly, $E[\max_{i \leq n} u(B_i)] = \sum_j u_j(F_j^n - F_{j-1}^n)$, where F_j is cumulative class probability and $F_0 = 0$: the difference of powers is the probability that the best of n draws has utility u_j . We normalise by the class utility range and integrate by the trapezoidal rule, dividing by the draw-count span $24 - 1 = 23$. Each draw selects a fixed mean campaign utility, not another adaptive assay within a campaign.

Elicitation protocols and replication unit. The sampled round estimates distributions from 8,088 replies over Claude Sonnet 4, DeepSeek V4 Flash, Gemini 2.5 Flash, GPT-4.1. The separately registered logit round uses Qwen2.5-0.5B/1.5B/3B, Llama-3.1-8B, and gpt-4.1-nano/mini and gpt-4o-mini: 1176 cells, 7 models in three families (Table 1). The 12 landscapes generate only 3 distinct prompts per arm: site count is the only landscape-dependent input. Thus choice-side uncertainty uses three prompt groups, not twelve independent replications. Re-clustered effects persist (blind renaming 0.770 [0.690, 0.811], aliasing 0.558 [0.458, 0.679], ordering 0.521 [0.380, 0.599]), but these small-cluster intervals are sensitivity analyses, not conventional 95% confidence intervals. Landscape pairs with identical prompts also expose between-batch drift: mean TV 0.173 [0.137, 0.226] versus 0.139 [0.114, 0.167] within a batch.

Instrument validity and eligibility. The 3 local models expose every identifier logit, with zero omitted permitted mass in 504 cells. The 4 hosted endpoints return at most 20 alternatives; probabilities are renormalised over returned permitted identifiers. A validation using 40 sampled draws on 8 landscapes passes for Qwen2.5-1.5B (0/8 disagreements within the multinomial floor), but fails for gpt-4.1-nano (4/8, $p = 0.00003$), with a further check finding hosted temperature scaling. Hosted distributions therefore supply sensitivity evidence. Worst-case allocation of omitted mass changes TV by median 0.00007 (maximum 0.079) without changing the primary verdicts; it does not resolve that sampling discrepancy. Of 1176 logit cells, 27 fail the 0.90 mass floor, all on ensembling arms; 0 were never collected. Canonical/raw contrasts are complete, while ensembling uses 645 eligible cells (Appendix P, Table 11, Figure 6).

Statistical protocol and robustness. The primary family, the endpoint (total-variation distance between selected behaviour-class distributions, judged against the replicate floor by a 10,000-relabelling permutation test [23]), the thresholds and the gates were fixed before the first model call. Intervals use the registered BCa cluster bootstrap [33]; using BCa rather than percentile endpoints changes no registered verdict. Leave-one-model-out analyses preserve 9 of 11 pooled verdicts; the two exceptions are labelled *model-set dependent*. Registered tests and subsequent re-analyses are distinguished in the results and Table 10 (Appendix N).

3 Results

3.1 The written menu is not the executable space

Of the 41 declared policies, 40–41 execute per landscape and collapse to 21–22 distinct campaigns. On every landscape, one class contains 19 entries, including upper-confidence-bound (UCB) and Thompson-sampling rules, exploration weights and stopping rules. At the standard configuration these options select the same variants in the same rounds. Four checks test stability beyond the original probes: held-out probe states split zero classes; within-class utility spread on *disjoint* evaluation seeds is 0.0000; ϵ -relaxation leaves the class count unchanged to $\epsilon = 0.1$; and adding query order to the signature splits zero classes (Appendix I). The biggest class has the same 19 members on all 12 landscapes. Changing 8 of 21 coordinates never changes execution at this base; we call these coordinates *inert*. Their redundancy is shared across this biological panel, rather than arising on one landscape. The audit costs a median 3.2 minutes of single-core compute and no additional assays. Of the coordinates, 8 of 21 are inert on every landscape and 0 on only some. Removing always-inert changes deletes 14 entries per landscape (40→26 on 9 landscapes; 41→27 on 3), 35% of the 40 executable entries at the former operating point. This deletes always-inert coordinate changes; full quotienting also merges other duplicates, yielding 21–22 classes. The ratio of entries to behaviours falls from $1.87\times$ to $1.22\times$. Each deleted entry executes as the default, preserving the audited behaviours unlike pruning that may discard a good configuration [29].

Independent grammars and AutoML replication. Two independently written design spaces show that the collapse is not specific to our curation, and supervised-learning pipeline selection establishes the same structure outside protein engineering. The curators had no access to our space, results or the optimum. One explicitly excluded anticipated duplicates; its grammar collapses fourfold less than ours, yet execution still finds 2 of its 18 coordinates inert; the second leaves 3 of 19 (Table 3). In supervised-learning pipeline selection, the same instrument maps 29 declared options to 12–18 behaviours. Crossing {our list, an independent curator’s list} with {our default, their default} and executing all four pools, same default, different list gives Jaccard 1.000—an *identical* reachable set—while same list, different default gives 0.020. In this comparison the default determines the reachable behaviours more strongly than the coordinate list. This yields two model-independent consequences: a reported design-space *size* overstates the alternatives available by $1.91\times$, and the syntactic-uniform null an evaluator gets by sampling the written menu differs from the behaviour-uniform null by 0.0050 fitness units. Protein-UQ reports uncertainty-aware optimisation often losing to uncertainty-free baselines [3]. Our single-model base supplies one possible mechanism: its uncertainty-aware rules execute as greedy mean. We have not established that this mechanism explains Protein-UQ’s results.

3.2 Equivalent rewritings change credit and model rankings

Equivalent rewritings change choice distributions. Over the primary family the median total-variation distance between the behaviour-class distribution chosen from the canonical menu and from a semantics-preserving rewriting is 0.417, against a same-prompt replicate floor of median 0.144 (maximum 0.292); 44 of 85 primary transformations are significant after Holm correction [10] within their cell, and every model family and every landscape contains one. The largest single effect is a permutation: on DHFR, reordering the menu moves Claude Sonnet 4’s distribution by $TV = 1.000$ ($p_{\text{Holm}} = 0.0003$) against a floor of 0.042, from *gaussian_process* in 23 of 24 calls to *expected_improvement*—the *worst* of the 21 available behaviours—in 24 of 24. The executable set is unchanged, and the effect extends beyond the top choice: Kendall’s τ between the induced rankings is 0.505 against 0.820 at the floor.

Credit verdicts and model rankings change. The three primary rewrites reverse 12/85 credit signs. Including four further orderings and shuffled aliases gives 38/193 reversals; secondary rewrites contribute 18/62 (56/255 pooled). The byte-identical replicate reverses 3/41 signs. These descriptive rates have different coverage; they are not a matched causal ratio. Figure 2 separates primary and secondary rewrites. The largest primary swing is 79.8% of the best-to-worst utility span, produced by reordering. Across the broader rewrite set, the top-ranked model changes in 15 of 26 rewritings and the sign of a two-system gap under 65 of 208 (31%). The confirmatory round repeats this on single-call distributions with only the arrangement varying: on 12 of 12 landscapes the top-ranked model is *not unique* and 70.3% of pair $\times$ landscape verdicts change sign, falling to 9 and 35.4%

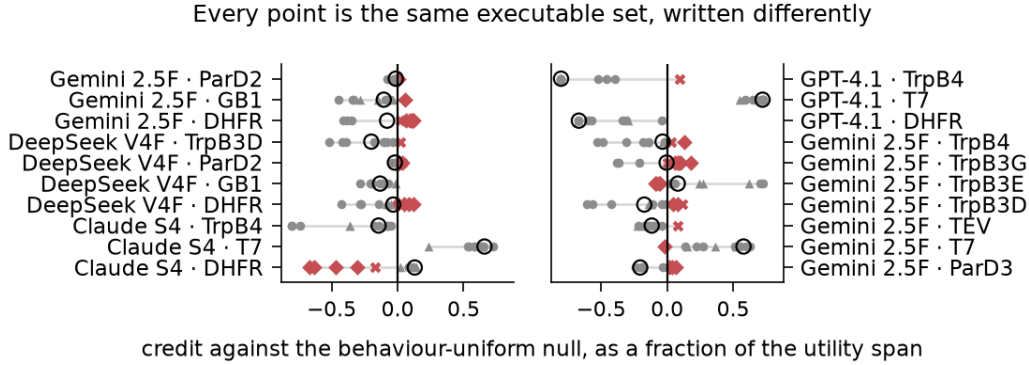


Figure 2: Credit under equivalent writings, estimated from sampled replies. Each row fixes model, landscape and executable set; all cells with a canonical estimate and at least six available rewrites are shown. Open circles: canonical. Grey dots: primary or additional order/alias rewrites; grey triangles: secondary rewrites. Red diamonds and crosses mark their respective sign reversals relative to canonical. Zero is the behaviour-uniform null; credit is divided by the best-to-worst utility span. Lines show the observed range, not confidence intervals. Crossings are point estimates; choice-distribution tests are reported separately.

under four-arrangement ensembling at the same number of draws (Appendix L). Both qualitative conclusions survive placing unobserved hosted mass adversarially (12/12, 69.9% against 70.3% as read), and both hold in the validated local family alone. Per landscape the standard deviation of credit across *menus* at a fixed model is 111% of that across *models* at a fixed menu: how the design space is written is an intervention of the same order as which model is used.

Held-out and cross-domain transport. The phenomenon holds on the nine landscapes held out of development (median primary-family TV = 0.458). A rule predicting credit reversals, fitted on the development slice, was sealed with per-cell predictions and a file hash before prospective testing; the result is directionally supported and not established — every measure points the predicted way, both sharp bars are missed, on 11 prospective cells (Appendix D). Under an unchanged instrument, the effect also appears in the AutoML space at 0.24–0.91 against a floor of 0.008, although verdict reversal is rare because the utility spread between pipelines is small. A separately sealed third grammar tests 24 orders of eight executable knapsack heuristics over 6 problem families. Both models clear the registered TV threshold on every family (median 0.291/0.771), while order spans both sides of the credit null on 6/6 families for Qwen2.5-0.5B and 1/6 for Qwen2.5-1.5B. Representation sensitivity persists across settings, while its magnitude depends on the grammar and model (Appendices D and U.4).

Two registered task-response probes pass in every model, but responsiveness alone does not establish useful policy selection (§4). Asked to identify equivalent policies, the best model recovers only 21% of the execution-verified equivalences (Appendix G).

3.3 Positional and textual preferences

Separating positional and textual effects. Choice probability falls with menu position ($\rho = -0.382 [-0.421, -0.334]$); controlling for utility gives $\rho = -0.375 [-0.421, -0.317]$ over 84 cells (Figure 3a). The curve differs from the U shape reported for long-context retrieval [34]: mass by third is 0.51/0.39/0.10, with least mass at the end. This is a short-context result (≈ 486 tokens). The bimodal curve confounds position with the option occupying it. Blinding every option’s text to a neutral code, leaving order and the executable set intact, separates the peaks. In a post hoc control motivated by this shape, primacy strengthens ($2.85\times$ to $3.14\times$ in the first octile) while the mid-menu peak collapses ($2.01\times$ to $0.54\times$). The early-entry preference persists without descriptions, whereas the middle peak depends on their wording. A registered test finds no evidence that doubling the 21-class quotient menu to the 41-entry declared menu steepens the slope ($-0.011 [-0.091, +0.033]$).

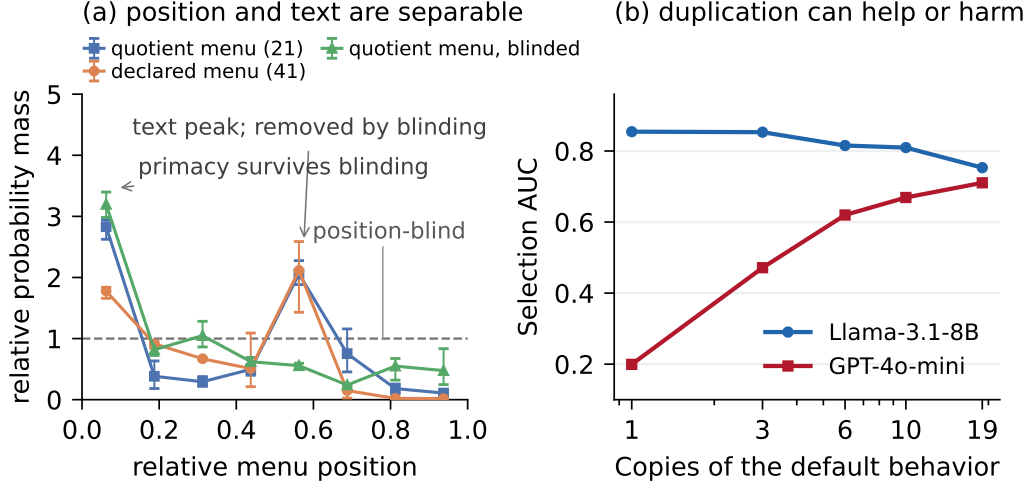


Figure 3: (a) Probability mass in each position bin relative to its share of menu entries (uniform selection = 1). Points average three prompt groups; bars span their means, not confidence intervals. Neutral codes strengthen the early-entry preference. (b) Duplicating the same default changes selection AUC in opposite directions for two preselected models; means over 12 landscapes, with slope intervals in §3.4. Selection AUC concerns repeated policy draws, not adaptive rounds within one campaign.

188 **Shared preferences and utility.** Do different models favour the same descriptions? For class c , let
189 $p_m(c)$ be model m 's selection probability and $z_m(c)$ its log probability centred across classes within
190 a landscape. We call the across-model mean $a(c)$ *shared attractiveness*: models agree at $\rho = +0.321$
191 $[+0.281, +0.359]$, and a computed from six models predicts the seventh in 7/7 held-out models
192 at $+0.27$ to $+0.79$. Attractiveness correlates with menu position ($\rho = -0.671 [-0.697, -0.640]$)
193 and jargon ($\rho = +0.603 [+0.560, +0.632]$). Alias-count and text-length associations are small
194 (-0.069 and -0.028). The jargon count and these post hoc descriptive comparisons are defined in
195 Appendix C. Shared attractiveness is negatively but imprecisely associated with utility, $\rho(a, u) =$
196 $-0.171 [-0.322, -0.009]$. A pre-registered control removing the no-op class—the strongly preferred
197 default—gives a mixed result: agreement ($+0.335 [+0.293, +0.377]$) and transfer (7/7) survive it, the
198 utility relation ($-0.166 [-0.315, +0.002]$) no longer excludes zero, so the utility result is restricted to
199 the full menu. The registered hazard score, intended to flag harmful arrangements, predicts discovery
200 no better than chance ($-0.131 [-0.385, +0.155]$, 20 cells) or the best option's position (-0.162
201 $[-0.438, +0.078]$). It therefore fails its predictive threshold.

202 A sealed paraphrase control holds identifiers and classes fixed while a separate local model (Qwen2.5-
203 0.5B) rewrites descriptions without seeing utilities. With Qwen2.5-1.5B full-vocabulary scoring,
204 median TV is 0.145 on 6 surfaces (6/6 above 0.05), but there are 0 top-class flips and credit is stable
205 on 6/6. Thus wording changes choices without necessarily changing credit. This uses one model
206 author and one scorer, not an independent human-author replication (Appendix U.1).

207 **Executed order-only intervention.** Because the positional weight is measurable, the reach of an
208 adversary who changes *only* the order is computable by the rearrangement inequality: it straddles the
209 matched null in 84/84 cells, moving expected utility across 0.52 of the span, and needs no access
210 to the target (Appendix P). Because real text may violate the ceiling's separability assumption, we
211 tested sealed best/worst orderings for 16 held-out cells built from six *other* models and hashed before
212 the first call. The pooled registered test does not meet its threshold: the direction is realised in 10
213 of 16 cells against a sealed bar of 12 ($p = 0.227$). The realised swing is $+0.248$ of the computed
214 one at the median, so the ceiling is an overestimate and the position-only model partially transports.
215 Despite this, order alone flips the sign of realised credit in 11 of 16 cells on menus verified entry
216 by entry to offer the same executable behaviours. Order-only manipulation is known for listwise
217 recommenders [30]. Here it changes a scientific benchmark's credit verdict despite execution-verified
218 equivalence of the options.

Model ordering underperforms behaviour-uniform selection. In the original seven-model logit round, canonical minus behaviour-uniform is negative in all seven and significant in three, using new calls and two additional families. This survives adversarial allocation of unseen hosted mass (0 of 7 models turn positive in the best case). We test whether rescaling confidence rescues this ordering by tempering $p_T \propto p^{1/T}$ (Appendix B). Within this temperature family, flattening recovers +0.661 [+0.383, +0.810] in AUC for gpt-4o-mini, significant in 5 of 7—but it recovers it by erasing the ordering: every curve converges to 0.871–0.874, the matched-uniform value 0.872. The tested curves favour the uniform limit, and $\rho(p, u)$ is negative in 6 of 7.

Choosing a temperature on the evaluation data can nevertheless create an apparent win. In-sample tuning makes gpt-4.1-mini at T^* appear to beat uniform (0.883 versus 0.872), but the preregistered leave-one-landscape-out excess over uniform is -0.007 [$-0.031, +0.018$], showing that the apparent gain is selection-induced. Out of sample, a temperature beats uniform in 0 of 7 models; re-scoring under 4 normalisations of the same executed utilities (as published, raw, rank, z) yields 0/0/0/0 of 7, respectively.

These temperature transformations do not establish that every decoder fails. They support behaviour-uniform selection over the tested distributions, consistent with Blankenstein et al. [24]; selection using additional scientific evidence remains outside this comparison.

3.4 Stable measurement and useful selection are different objectives

Behavioural quotient normalisation (BQN) executes every candidate on frozen probes and presents one entry per behaviour class in a fixed order. It changes the chooser’s task: the model no longer has to recognise aliases. By construction it is invariant under the primary family and cannot discard the best class; empirically it removes a median 62.0% of the credit spread, leaving chiefly the secondary family’s blind-renaming effect.

Representation invariance score. Define $\text{RIS} = 1 - \text{clip}[(\text{med}_T \text{TV}(P_M(\cdot \mid D), P_M(\cdot \mid T(D))) - \phi)/(1 - \phi), 0, 1]$, for replicate floor $\phi < 1$. The tested family contains the three primary rewrites and available additional orderings, with at least three eligible rewrites per cell. $\text{RIS} = 1$ means the median distance does not exceed the measured floor; it does not certify invariance. Distances (0, 0, 1) with zero floor give $\text{RIS} = 1$ despite one maximal violation. Across 28 sampled cells, family means span 0.37–0.94. Report the tested transformations and upper-tail distances alongside RIS and utility; a stable chooser can still select poorly.

No detected pooled utility gain. BQN improves all three registered coverage endpoints, while selection AUC changes by +0.010 [$-0.046, +0.086$] and does not reject the Holm-corrected null. This interval permits both harm and benefit; it is not an equivalence test. Execution-pruning likewise gives no detected improvement (3/12, $p = 0.176$). The following intervention shows why a stable representation need not yield better selection.

Multiplicity causally changes selection. We selected one strong and one weak chooser by canonical AUC before making any dose call, then wrote the same medium-utility behaviour $k \in \{1, 3, 6, 10, 19\}$ times while holding the executable class set fixed. Selection AUC falls with $\log k$ for llama-3.1-8b-instruct (slope -0.033 [$-0.079, -0.009$]) but rises for gpt-4o-mini (+0.178 [+0.101, +0.243]); the registered crossing test passes. These dose distributions are hosted estimates, subject to §2’s validation boundary. From $k = 1$ to 19, the former moves from 0.854 to 0.753 and the latter from 0.199 to 0.710. Utilities, probes, prompt template and call budget stay fixed. Duplication attracts these two choosers toward the default with opposite utility effects (Figure 3b). This does not establish a general sign law: the sealed prediction and AutoML transfer fail (§4). Quotienting can remove a useful prior. Appendix K reports within-model sign flips.

Explaining versus predicting duplication. Does duplication merely attract probability to the default, or also change preferences among the other policies? A retrospective test supplies the observed default probability and keeps the remaining probabilities in their original ratios. Across 48 protein dose contrasts per model, its selection-AUC error is 0.003 for GPT-4o-mini but 0.062 for Llama and 0.112 for GPT-4.1-nano: the latter models also redistribute non-default mass. A predictor that instead multiplies the baseline default weight by k has error 0.424 for GPT-4o-mini, essentially the 0.425 no-change error. Thus a close conditional fit does not establish a predictive law.

These descriptive errors reuse the hosted dose estimates. A fresh, sealed challenge collects 84 full-vocabulary local distributions on three AutoML datasets, two default placements and two Qwen scales. Both scales fail the fixed approximation and prediction criteria; for 1.5B, duplication even lowers default mass in 18/24 contrasts. Numerical repeats agree exactly. These are bounded counterexamples, not evidence that duplication destroyed an initially competent selector (Appendix X).

Averaging over menu arrangements. Representation ensembling averages the measured class distributions across arrangements before computing repeated-selection AUC. Unlike one-draw expected utility, this best-of-several endpoint is nonlinear. Averaging reduces dependence on the curator’s choice but does not dominate the best arrangement: it beats the mean single arrangement in 7/7 models and the registered canonical arrangement in only 3. At matched total calls, one arrangement reverses the all-arrangement verdict 11.7% of the time; four cut this to 4.6% and eight to 1.4%. Four arrangements are a supported trade-off in this panel (Appendix U.2). Neither repair establishes a pooled advantage over behaviour-uniform (§3.3); Appendix L gives full estimates.

4 Discussion

Selection with prior experimental evidence. When models see only site count and method descriptions, relabelling changes choices more than substituting the task; the registered prediction of redundancy effects nevertheless fails prospectively and in AutoML. We next supplied prior experimental outcomes to two local Qwen scales. Evidence from other landscape groups did not let even the reference selector pass the prespecified competence criterion. Using 378 fresh pilot campaigns on six landscapes, a reference that chooses the best pilot mean does pass. For a model to pass, the lower bootstrap bound on its credit over behaviour-uniform selection must be positive. The 1.5B gain from adding evidence is $+0.062 [-0.011, +0.111]$; neither model passes. These tests leave the value of evidence-informed selection inconclusive (Appendices D and W).

Related work. Discovery agents search human-supplied program skeletons or modules [7, 6, 8, 14, 15, 2]; AutoML and neural architecture search establish the importance of search-space design [4, 5, 11]. Order, identity and format sensitivity are also known [16, 18, 17, 19]. Tool-menu and harness studies vary the action set or execution process [31, 32]; ToolTweak steers selection through tool names and descriptions [35]. Our contrast holds execution-verified opportunities fixed and measures discovery credit. It extends efficiency uses of execution equivalence [12, 13, 27] to evaluation. Related repairs filter equivalent tools and sample uniformly [24] or temperature-correct preferences [20]. Our results distinguish their measurement benefit from discovery utility (Appendix E).

5 Limitations and conclusion

Equivalent menus can change model choices and scientific credit while preserving the executable opportunity set. Duplicating one default also changes utility in opposite directions for two selected models. An evaluation should therefore report behaviour-level credit, transformation-specific sensitivity and utility separately; stable measurement does not guarantee useful selection.

The evidence is probe-relative. The primary protein selector sees site count and method descriptions, not observed sequences or fitnesses. The additional evidence-informed qualification is inconclusive; free-form generation remains outside scope. Two protein grammars were independently curated. AutoML supplies an independently sourced non-protein grammar; the knapsack transport remains researcher-constructed and uses two scales from one family. Failed prospective predictions limit general laws. These boundaries leave a concrete evaluation requirement: verify what the options execute before attributing a menu-conditioned advantage to the model.

AI use, ethics, and reproducibility

Language models supported research, code, analysis and writing under human direction; authors take responsibility. The appendix records protocols and amendments. Experiments use published landscapes, with no human subjects, personal data or wet-lab work.

References

- [1] J. Bradshaw et al. BioDiscoveryAgent: An AI agent for genetic disease target discovery. In *ICLR*, 2025.
- [2] J. Kim and P. A. Romero. Benchmarking and behavioral characterization of LLM agents for protein design. *bioRxiv* 2026.05.06.723381, 2026. <https://www.biorxiv.org/content/10.64898/2026.05.06.723381v1>.
- [3] K. P. Greenman, A. P. Amini, and K. K. Yang. Benchmarking uncertainty quantification for protein engineering. *PLOS Computational Biology*, 21(1):e1012639, 2025.
- [4] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.
- [5] L. Li and A. Talwalkar. Random search and reproducibility for neural architecture search. In *Uncertainty in Artificial Intelligence (UAI)*, 2019.
- [6] A. Novikov et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. 2025.
- [7] B. Romera-Paredes et al. Mathematical discoveries from program search with large language models. *Nature*, 2024.
- [8] N. van Stein and T. Bäck. LLaMEA: A large language model evolutionary algorithm for automatically generating metaheuristics. *IEEE Transactions on Evolutionary Computation*, 29(2):331–345, 2025.
- [9] R. T. Lange, Y. Imajuku, and E. Cetin. ShinkaEvolve: Towards open-ended and sample-efficient program evolution. In *ICLR*, 2026.
- [10] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 1979.
- [11] A. Yang, P. M. Esperança, and F. M. Carlucci. NAS evaluation is frustratingly hard. In *ICLR*, 2020.
- [12] F. Hutter, H. H. Hoos, K. Leyton-Brown, and T. Stützle. ParamILS: An automatic algorithm configuration framework. *JAIR*, 36:267–306, 2009.
- [13] H. Peleg and N. Polikarpova. Perfect is the enemy of good: Best-effort program synthesis. In *OOPSLA*, 2020.
- [14] C. Lu, C. Lu, R. T. Lange, Y. Yamada, S. Hu, J. Foerster, D. Ha, and J. Clune. Towards end-to-end automation of AI research. *Nature*, 651(8107):914–919, 2026.
- [15] J. Gottweis et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026.
- [16] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang. Large language models are not robust multiple choice selectors. In *ICLR*, 2024.
- [17] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying language models’ sensitivity to spurious features in prompt design, or: How I learned to start worrying about prompt formatting. In *ICLR*, 2024.
- [18] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *ACL*, pages 8086–8098, 2022.
- [19] J. Ye, Y. Wang, Y. Huang, D. Chen, Q. Zhang, N. Moniz, T. Gao, W. Geyer, C. Huang, P.-Y. Chen, N. V. Chawla, and X. Zhang. Justice or prejudice? Quantifying biases in LLM-as-a-judge. *arXiv:2410.02736*, 2024.
- [20] H. Yin, S. Vardi, and V. Choudhary. Fragile preferences: A deep dive into order effects in large language models. *PNAS Nexus*, 5(8):pgag246, 2026.

- [21] Y. Zhu et al. Establishing best practices in building rigorous agentic benchmarks. In *NeurIPS Datasets and Benchmarks Track*, 2025.
- [22] J. Wang, F. Bianchi, S. Zhu, F. Nie, Y. Kwon, B. Dhingra, and J. Zou. Automated benchmark auditing for AI agents and large language models. arXiv:2605.26079, 2026.
- [23] H. Helm and C. E. Priebe. When prompt perturbations break your A/B test: A valid statistical test for generative surveying. arXiv:2605.27463, 2026.
- [24] T. Blankenstein, J. Yu, Z. Li, V. Plachouras, S. Sengupta, P. Torr, Y. Gal, A. Paren, and A. Bibi. BiasBusters: Uncovering and mitigating tool selection bias in large language models. In *ICLR*, 2026. arXiv:2510.00307.
- [25] S. Weng, Y. Feng, and X. Xie. Beyond accuracy: Policy invariance as a reliability test for LLM safety judges. arXiv:2605.06161, 2026.
- [26] I. de Zarzà, J. de Curtò, J. Cabot, P. Manzoni, and C. T. Calafate. Semantic invariance in agentic AI. arXiv:2603.13173, 2026.
- [27] W. Huang, Q. Zeng, Q. Chen, Z. Guo, Y. Sun, C. Yang, S. Ouyang, J. Gesi, F. Wu, J. Zhang, H. Chen, B. Liu, X. Tang, and C. Wu. Latent action reparameterization for efficient agent inference. arXiv:2605.18597, 2026.
- [28] A. M. Bean et al. Measuring what matters: Construct validity in large language model benchmarks. In *NeurIPS Datasets and Benchmarks Track*, 2025.
- [29] M. Wistuba, N. Schilling, and L. Schmidt-Thieme. Hyperparameter search space pruning—a new component for sequential model-based hyperparameter optimization. In *ECML PKDD*, 2015.
- [30] G. Zhang, J. Cheng, and H. Chen. Ranked by position: Order sensitivity as an exploitable attack surface in LLM listwise recommenders. arXiv:2607.24869, 2026.
- [31] R. Suresh Babu and L. Ganesh Iyer. ToolMenuBench: Benchmarking tool-menu filtering strategies for reliable and efficient LLM agents. arXiv:2606.15508, 2026.
- [32] H. Yi and X. Song. Measuring harness-induced belief divergence in multi-step LLM agents. arXiv:2607.04528, 2026.
- [33] B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.
- [34] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.
- [35] J. Sneh, R. Yan, J. Yu, P. Torr, Y. Gal, S. Sengupta, E. Sommerlade, A. Paren, and A. Bibi. ToolTweak: An attack on tool selection in LLM-based agents. arXiv:2510.02554, 2025.
- [36] W. Bolton and P. Torr. Adversarial fast-moving real-world domains as test beds for benchmarking AI scientist capabilities. arXiv:2608.03569, 2026.

Reproducibility statement

Headline estimates are generated from retained results, with primary, extended-primary and secondary credit contrasts reported separately. This document records the prespecification summaries, dated amendments, sealed-prediction hashes, per-cell coverage, execution checks and limitations. A fully anonymised artifact containing the executed quotients, written spaces, analysis code, all 8,088 cached selection replies, and a clean replay will be released with the archival version. The prior manuscript, data manifest, and environment lock are preserved separately and were not modified by experiments reported here.

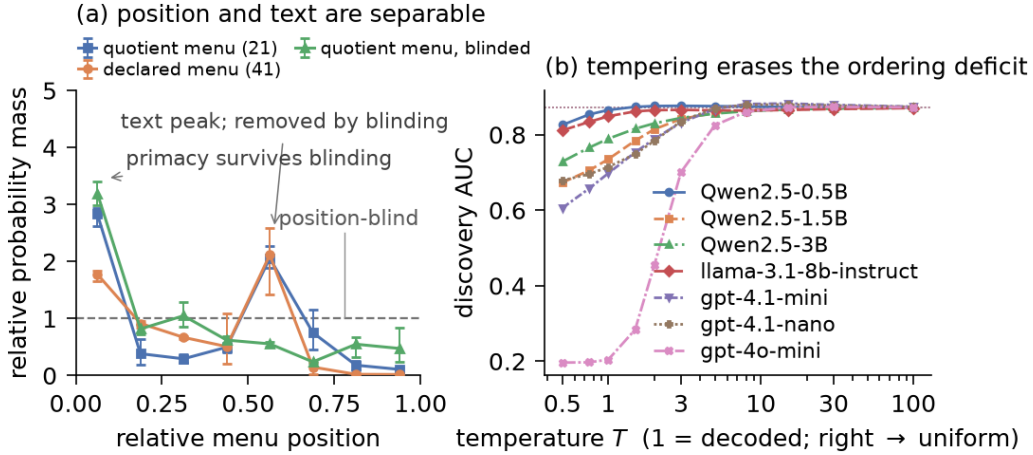


Figure 4: Position controls and temperature sweeps. The tested temperature family approaches behaviour-uniform; this does not exclude better decoders using other information.

LLM-use disclosure

Language models are both the object of study and instruments of production. Under human direction, agents supported literature mapping, hypothesis generation, protocol drafting, code and analysis, internal review, and prose editing. The authors selected the questions, approved sealed protocols and decision thresholds, audited the artefacts and claims, and take responsibility for the paper.

A Full experimental setup

Throughout, the campaign budget means 96 adaptive assays after 24 initial observations: 120 variants are observed in total. The twelve landscapes are the complete-landscape subset of the census: every variant of the mutated sites has a measured fitness, so a policy can be executed exactly rather than simulated. Landscapes whose optimum is the wild type are excluded, since there is nothing to optimise. The same declared design space, campaign executor, classical default and assay budget are used across the menu comparisons.

B Tempering and serial-position analyses

Figure 3 in the main paper shows position and duplication. Figure 4 retains the temperature curves; this section gives the numbers and registered status.

Tempering (R15). Grid $T \in [0.5, 100]$ on each model’s own single-call distribution. Every curve rises to that model’s matched-uniform value and stops there. PROSPECTIVE: the leave-one-landscape-out estimate was registered as the verdict *before* the in-sample table was read, precisely because choosing T^* on the data that scores it inflates the comparison. In sample gpt-4.1-mini at T^* scores 0.883 against uniform’s 0.872, which reads as a win; held out it returns $-0.007 [-0.031, +0.018]$. Held out, 0 of 7 models beat uniform at any temperature; $\rho(p, u)$ is negative in 6 of 7.

Serial position (R20). RE-ANALYSIS, sealed before computation (PRESPECIFICATION_R20_SERIAL_POSITION.md). Positions are recoverable on exactly the two arms that assign identifiers by `canonical_g`, and that was asserted on every cell before scoring. H1a $\rho = -0.382 [-0.421, -0.334]$; H1b, utility partialled out, $-0.375 [-0.421, -0.317]$, over 84 cells. H2 registered both directions in advance and finds an early-entry preference without recency: mass by third 0.51/0.39/0.10. H3—does doubling the menu to 40 options steepen the slope—FAILS ($-0.011 [-0.091, +0.033]$), so the pruning recommendation of \$3.1 rests on redundancy alone. The blinding control is POST HOC: it was chosen after the curve turned out to be

bimodal, and it is what separates the two channels ($2.85 \times$ to $3.14 \times$ first octile; mid-menu $2.01 \times$ to $0.54 \times$).

C Shared attractiveness and descriptive features

For each landscape, we floor class probabilities at 10^{-6} , take logs, centre across classes within each model and average across available models (at least four). This yields the shared-attractiveness score used in Section 3.3. We then compute Spearman correlations within each landscape and average them across the 12 landscape conditions, which share three task-prompt groups.

The jargon score counts which of 22 substrings occur in the lowercased canonical description: gaussian, process, ensemble, bayes, posterior, thompson, ucb, acquisition, rollout, lookahead, adaptive, expected, improvement, boosted, pca, latent, hierarchy, calibrat, surrogate, uncertainty, entropy, diverse. Each contributes at most one point; the count is not length-normalised and uses no fitness outcomes. Length is the canonical description’s character count; alias count is the number of entries for that class in the alias-expanded menu. Mean correlations are $+0.603$ for jargon, -0.671 for position, -0.069 for alias count and -0.028 for length. These post hoc associations do not isolate causal feature effects. The alias-count correlation compares different classes; the duplication intervention instead varies multiplicity for one fixed class.

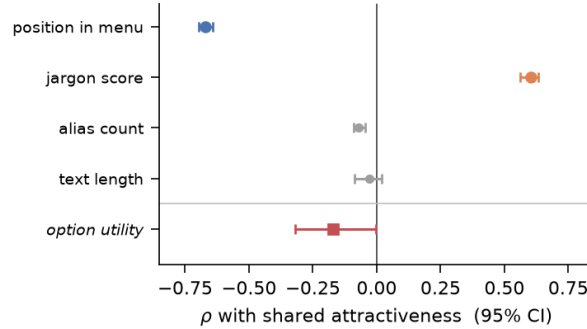


Figure 5: Shared attractiveness is calculated from model probabilities, then compared with description features. Models agree at $\rho = +0.321$ [$+0.281, +0.359$], with transfer to 7/7 held-out models. Position and jargon have the largest descriptive associations. The utility association is negative, modest and imprecise (§3.3); feature correlations do not establish a causal decomposition.

D Held-out and cross-domain validation

Developed on three landscapes chosen by ruggedness alone and extended to the other nine through the identical pipeline, the phenomenon holds out of development: median primary-family TV = 0.458, with 16 of 17 cells Holm-significant over 9 landscapes. A regime law the development slice suggested—that whether non-invariance corrupts the *verdict* depends on the margin $z = |C_{\text{canonical}}|/\text{span}$ —was fitted on those nine cells only and **sealed with per-cell predictions and a file hash** before the menus it predicts had been called. The registered thresholds are *not* met. Reversal is more common in the high-risk stratum (26% against 15%), but the rate ratio is 1.69 where P1 demanded 3 (one-sided Fisher $p = 0.23$), and the rank correlation reaches $\rho = +0.41$ against a threshold of 0.4 but at $p = 0.10$. The two weaker predictions do hold out of sample: the sealed probabilities beat a constant base rate in Brier score (0.157 against 0.159), and the model-specific margin outranks the landscape-level crowding statistic it was registered against (AUC 0.64 against 0.54). We therefore report the regime law as **directionally supported and not established**: every measure points the predicted way and the two sharp bars are missed, on the 11 prospective cells the budget allowed rather than the larger set the design called for. Registering the rule before the data is what makes it possible to say that instead of reporting the pattern as a finding. In the AutoML space sensitivity to an equivalent rearrangement is 0.24–0.91 against a floor of 0.008, though verdict reversal is rare there (2/66 rewritings) because the utility spread between pipelines is small: **the phenomenon generalises, its magnitudes are properties of a grammar**.

Table 2: Positioning against prior work. Prior work rewrites a *description* and measures whether the *answer* moves; equivalence is argued from semantics. We rewrite a *design space*, prove the executable opportunity set is unchanged by running it, and measure the *credit an evaluation assigns*.

work	what is rewritten	equivalence established by	endpoint	headline
Sclar et al. [17]	prompt format: separators, casing, spacing	construction	task accuracy	up to 76-point spread
Zheng et al. [16]	option IDs and their order	same options permuted	answer accuracy	selection bias from token priors
Yin et al. [20]	order of the items compared	same items	stated preference	quality-dependent primacy/recency
Blankenstein et al. [24]	tool list: order, descriptions, provider	<i>declared</i> functional equivalence	which tool is called	provider fixation; position bias
Weng et al. [25]	the judge’s rubric	certified-equivalent rewrite	safety verdict	up to 9.1% of verdicts flip
de Zarzà et al. [26]	the problem statement	eight metamorphic transforms	response invariance	best model invariant 79.6%
this paper	the executable design space: order, aliases, inert-coordinate multiplicity	execution of every option on frozen probe states	discovery credit vs. a matched null; the leaderboard	12/85 credit reverse; non-unique on 12/12

E Extended related work and positioning

F Confirmatory elicitation protocol

Every option identifier is a single character drawn from [A–Za–z0–9], so a choice distribution is read from one call’s `top_logprobs` rather than estimated from hundreds of samples; median captured probability mass over permitted identifiers is ≥ 0.976 in every family. *What that read is differs by family, and §2 and Appendix P say so before any result uses it:* the three Qwen rungs were run locally from the full vocabulary logits, where the distribution over identifiers is exact and the replicate floor is identically zero; the four hosted models return at most the top 20 alternatives, of which the permitted identifiers are kept and renormalised, so those are truncated distributions. One family, `deepseek-v4-flash`, is *excluded with the reason:* the router serves it from providers that return no per-token alternatives for these prompts. A measurement fact the literature should record: at temperature 0, sending the byte-identical prompt twice, the two OpenAI endpoints return *different* distributions—replicate floors of 0.080 and 0.161 in total variation—while the open-weight models and the Llama endpoint are exactly reproducible. Every OpenAI figure in the paper is reported against its own measured floor. Two later arms are labelled amendments and excluded from the confirmatory endpoint.

G Rejected alternative explanations

It is not a misunderstanding the model can be talked out of: telling it in the same prompt that repetition carries no information, that order carries none, or both, does not move the distribution back (mean change in distance from the canonical distribution +0.013 over 18 warned arms, helping in 6 of 18, one-sided $p = 0.76$). And it is not redundancy the model could have priced in—asked to partition the declared menu into sets that execute identically, the best frontier reasoning model, Claude Opus 5, recovers 21% of the true equivalences at precision 0.70, and non-reasoning families are effectively blind (Figure 8).

H Complete written design spaces

The **primary family** is: *order permuted*, five independent orderings; *alias expanded*, every verified alias in a class written out as its own entry, plus a shuffled variant that keeps the multiplicity but destroys the grouping; and *dead inflated*, entries added that differ only in a coordinate execution

Table 3: Every design space we executed. **declared**: written options in the space (mean over landscapes). **delivered**: distinct campaigns after exact behavioural quotienting. **largest**: options in the biggest equivalence class. The collapse is not specific to our curation, our task or our domain; a curator who set out to prevent it reduced it fourfold and did not remove it.

design space	executed on	declared	delivered	redundancy	largest class
this paper’s grammar	12 protein landscapes	41	21–22	$1.91\times$	19
independent grammar B	10 protein landscapes	34	28–29	$1.19\times$	3
independent grammar C	9 protein landscapes	37	27–28	$1.34\times$	7
AutoML pipeline space	3 tabular tasks	29	12–18	$1.99\times$	15

proves inert here. Invariance is normatively required across it because nothing an agent could legitimately condition on changes. The **secondary family** is reported separately because a reader could legitimately condition on what it changes: *default anchored*, where the literature-standard configuration is labelled conventional; *blind renamed*, where coordinate values are replaced by neutral codes; and *verbosity padded*, the same entries in longer prose, which separates prompt length from entry multiplicity. Every constructed space is admitted only after *verify* confirms by execution that its quotient equals the canonical quotient—textual plausibility is not sufficient—and permutation arms are additionally checked for entry-multiset identity. Two later arms are labelled amendments and excluded from the confirmatory endpoint.

Quotienting a 41-option menu costs a median of 3.2 minutes of single-core compute and no assays. Averaging over permutations instead—the obvious alternative to quotienting—needs up to 129 independently ordered menus to reach the same credit variance, at $129\times$ the API cost and with no protection against alias multiplicity. Per-cell testing is underpowered at realistic budgets: with 24 calls on each of two arms a permutation test on the behaviour distributions rejects in 67% of rewriting comparisons and 41% at eight calls, while its false-positive rate against the byte-identical replicate is 0.7%.

I Experimental-object validity (Gate 0)

Four checks run before any model call. **Fidelity**: the instrumented rollout that records selections must reproduce the frozen harness’s own fitness trace exactly, asserted on every candidate (488 checks, zero failures). **Probe-order stability**: the partition must not change when the probe states are visited in the reverse order. **Held-out probes**: four additional probe states are added and the number of classes that split is reported (zero). **Order sensitivity**: adding within-round query order to the signature splits zero classes. Every constructed written space is then verified by execution to have the same quotient as the canonical space, and permutation arms are checked for entry-multiset identity; there were no failures.

J Tolerance sensitivity of the quotient

Exact equality is the primary definition. Relaxing it to ϵ -equivalence—mean Jaccard distance between the selected variant sets at most ϵ , single-linkage—leaves the class count unchanged from $\epsilon = 0$ through $\epsilon = 0.3$ on every landscape tested, and the minimum between-class distance at the exact quotient is far from zero. The partition is therefore not a knife-edge property of demanding exact equality; distinct behaviours select substantially different variants.

K Alias multiplicity and model selection

The declared menu writes the same behaviour class nineteen times. Table 6 gives the probability each model places on that one class, on the quotient menu and on the declared menu; it is the per-model detail behind C2’s no-op endpoint in §3.4.

Table 4: The declared design space, executed. **written**: declared options; **behaviours**: distinct campaigns after exact behavioural quotienting on frozen probe states; **largest**: options in the biggest equivalence class; **split**: classes split by four held-out probe states; Δu : largest within-class utility spread on the *disjoint* evaluation seeds. The first three landscapes are the development slice; the remaining nine were built afterwards with the identical pipeline.

landscape	sites	rugged.	written	behaviours	redund.	largest	invalid	split	Δu
DHFR	3	0.0000	41	21	1.95×	19	1	0	0.0000
TrpB4	4	0.0060	41	22	1.86×	19	1	0	0.0000
T7	3	0.0955	41	22	1.86×	19	0	0	0.0000
GB1	4	0.0010	41	22	1.86×	19	1	0	0.0000
ParD2	3	0.0010	41	21	1.95×	19	1	0	0.0000
ParD3	3	0.0010	41	21	1.95×	19	1	0	0.0000
TEV	4	0.0070	41	22	1.86×	19	1	0	0.0000
TrpB3D	3	0.0025	41	21	1.95×	19	1	0	0.0000
TrpB3E	3	0.0120	41	22	1.86×	19	0	0	0.0000
TrpB3G	3	0.0115	41	21	1.95×	19	1	0	0.0000
TrpB3H	3	0.0140	41	22	1.86×	19	0	0	0.0000
TrpB3I	3	0.0020	41	21	1.95×	19	1	0	0.0000

Table 5: Classes surviving as ϵ grows (column headers give ϵ). $\epsilon = 0$ is the exact quotient used everywhere else. **min. dist.**: smallest between-class distance at the exact quotient.

landscape	0	0.05	0.1	0.2	0.3	0.5	0.75	min. dist.	median dist.
DHFR	21	21	21	21	20	16	10	0.248	0.815
T7	22	22	22	21	17	12	7	0.163	0.691
TrpB4	22	22	22	22	22	22	18	0.684	0.937
ParD2	21	21	21	21	21	18	9	0.322	0.818
TrpB3H	22	22	22	22	22	22	16	0.654	0.932

Table 6: Probability mass on the single behaviour class the declared menu writes out nineteen times.

model	quotient menu	declared menu
Qwen2.5-0.5B	0.018	0.874
Qwen2.5-1.5B	0.139	0.989
Qwen2.5-3B	0.005	0.022
Llama-3.1-8B	0.012	0.533
gpt-4.1-nano	0.000	0.937
gpt-4.1-mini	0.000	0.112
gpt-4o-mini	0.000	0.845

535 L Effects and limits of representation ensembling

536 Let G be the group generated by permuting menu order and reassigning option identifiers. Every
537 element of G leaves the executable behaviour of every entry unchanged. Averaging over the complete
538 finite group is invariant to the starting arrangement. Our empirical average $\bar{p} = K^{-1} \sum_k p_{g_k}$ uses
539 only K sampled arrangements and can retain representation dependence. The utility endpoint first
540 draws policies independently from this averaged distribution, then retains the best fixed class utility.
541 Its AUC need not equal the mean AUC of individual arrangements, because best-of-several selection
542 is nonlinear. For one draw, expected utility of the average does equal average expected utility. We
543 registered four hypotheses about this finite approximation (Table 9).

544 **H1 partially passes.** The spread of discovery AUC across arrangements shrinks significantly in
545 every model and meets the registered 50% bar in 4 of 7; leave-one-model-out marks that bar as a
546 threshold artefact—the pooled estimate is 0.501 against a bar of 0.50, so 4 deletions move it across.
547 The reduction itself is robustly positive; we do not move the bar.

Table 7: Discovery AUC against the number of times one behaviour is written into the menu. The class set is identical at every level; only multiplicity changes. The slopes have opposite signs.

model	$k=1$	$k=3$	$k=6$	$k=10$	$k=19$	slope in $\log k$
Llama-3.1-8B	0.854	0.853	0.815	0.810	0.753	-0.033 [-0.079, -0.009]
GPT-4o-mini	0.199	0.471	0.620	0.669	0.710	+0.178 [+0.101, +0.243]

Table 8: Sign flips *inside* a model — the cases a two-model comparison cannot produce.

model	landscape	A_0	A_m	$A_0 - A_m$	slope
llama-3.1-8b-instruct	TrpB3H	0.705	0.658	+0.046	+0.010
llama-3.1-8b-instruct	TrpB4	0.960	0.921	+0.039	+0.015
llama-3.1-8b-instruct	TrpB3E	0.175	0.886	-0.710	-0.185
gpt-4o-mini	T7	0.258	1.000	-0.742	-0.024

548 **Mean-arrangement comparison (H2).** The finite ensemble beats the expected single-arrangement
549 AUC in every model under both clusterings.

550 **Canonical-arrangement comparison (H3).** The registered condition was that the ensemble beat
551 the *canonical* arrangement in each family; it does so in 3 of 7. We state the weaker claim the data
552 support: *ensembling reduces dependence on the curator’s arrangement without a detected pooled*
553 *discovery loss, and does not beat the best single arrangement.* We checked why rather than assuming.
554 The canonical arrangement is not a neutral draw from G —for some models. Its percentile within G is
555 0.65, 0.57, 0.55, 0.53 and, for Qwen2.5-1.5B, 0.24, where the declared arrangement is significantly
556 *worse* than a random one. Which arrangement is lucky is itself model-specific, so a curator cannot
557 assume the same arrangement will favour every model. For gpt-4o-mini, the declared menu sits -0.667
558 below matched uniform. Ensembling helps in this poor-arrangement case, but does not consistently
559 improve an already favourable arrangement. Since a curator cannot tell in advance which they have
560 written, the case for ensembling is insurance, not gain. One distinction must not be blurred: the
561 *pooled* form of H3 passes and is robust, but the **registered** condition was per family and it fails. The
562 registered verdict stands; the pooled statistic is reported beside it as a different quantity, not as a
563 rescue.

Table 9: Representation ensembling, $K = 4$, discovery AUC. Landscape-clustered 95% intervals over twelve landscapes. **H2 passes in every model; H3 passes in 3 of 7.**

model	H1 spread reduction	H2 – mean single	H3 – canonical	canonical – uniform
Qwen2.5-0.5B	0.470 [0.411, 0.560]	+0.031 [+0.020, +0.049]	-0.008 [-0.062, +0.019]	-0.006 [-0.032, +0.047]
Qwen2.5-1.5B	0.407 [0.355, 0.504]	+0.057 [+0.036, +0.109]	+0.126 [+0.084, +0.174]	-0.136 [-0.229, -0.077]
Qwen2.5-3B	0.619 [0.540, 0.691]	+0.088 [+0.060, +0.125]	+0.029 [-0.013, +0.131]	-0.082 [-0.118, -0.058]
Llama-3.1-8B	0.528 [0.451, 0.600]	+0.040 [+0.027, +0.055]	-0.033 [-0.115, +0.018]	-0.023 [-0.088, +0.064]
gpt-4.1-nano	0.616 [0.461, 0.724]	+0.107 [+0.070, +0.170]	+0.089 [+0.006, +0.229]	-0.157 [-0.261, -0.074]
gpt-4.1-mini	0.647 [0.567, 0.729]	+0.100 [+0.067, +0.132]	+0.030 [-0.007, +0.062]	-0.174 [-0.301, -0.043]
gpt-4o-mini	0.218 [0.140, 0.319]	+0.190 [+0.120, +0.248]	+0.414 [+0.251, +0.557]	-0.667 [-0.816, -0.388]

564 M Model capability, invariance, and competence

565 The registered rank correlation is evaluable and it **passes**. A disk stop rule had left the open-weight
566 ladder at two rungs; when disk freed, the third registered rung was run with the family, the endpoint
567 and the test unchanged. Sensitivity to $g \sim G$ is 0.645 at 0.49B parameters, 0.861 at 1.54B and 0.771

at 3.09B; Spearman $\rho = +0.50$, landscape-clustered CI $[+0.50, +0.50]$. The registered criterion was $\rho > -0.5$ or a CI including zero, so **R2-H1 passes: sensitivity does not fall with scale in this family**. Two caveats belong beside that verdict. With 3 rungs a Spearman can take only four values, so the criterion is decided entirely by whether the largest rung is the least sensitive—it is not, the smallest is. And the CI is degenerate because across 4,000 landscape resamples the rung ordering never changes, which is a statement about stability rather than precision. The ladder is non-monotone, so the defensible reading is the registered one: *no evidence of a fall*, not a rise.

Within what can still be measured, the registered *competence* hypothesis *fails*: expected utility of the choice does not rise with scale, and falls in both within-family pairs. Across all 7 measured models the cross-model relationship agrees and is not significant ($\rho = +0.46$, permutation $p = 0.31$, $n = 7$); the point estimate is positive and we claim nothing more from 7 points. The relationship that *is* significant is the confound, and it is the more important finding. Apparent competence tracks the *entropy* of the choice distribution almost perfectly ($\rho = +0.71$, $p = 0.091$). 5 of 7 models sit better-matched behaviour-uniform and none clears it by a meaningful margin, so what separates the better-scoring models is how close they come to not selecting at all. The extreme case makes the point: gpt-4o-mini is the *least* representation-sensitive model here (0.240), is essentially deterministic (0.02 bits), and is by a wide margin the worst chooser (0.197 against a matched-uniform 0.613). **An invariance score used on its own would rank it first**. Representation invariance is a necessary property of a trustworthy interface, never a sufficient one.

N Model-set robustness of pooled verdicts

Every pooled contrast macro-averages across models, and one verdict in an earlier draft reversed when the model set grew. We therefore recompute each verdict on every subset that deletes one model. **9 of 11 verdicts are unchanged under any single deletion**. The two that are not have different causes and are separated rather than reported as one fragility rate: the spread-reduction bar is a threshold artefact (Appendix L), and C2’s primary is genuinely model-set dependent, on exactly the model whose mechanism §3.3 identifies—a dependency with a cause, not diffuse noise. Both are labelled *model-set dependent* wherever they appear.

O Registered generation-versus-selection test

The most attractive claim in the earlier draft was that the model generates better policies than it selects. That result generated proposals for one landscape and executed them on eleven others. We registered a target-conditioned replication—fresh generation for each landscape, 155 executed policies over 6 held-out landscapes, each paired with a random policy at *matched coordinate reach*—and it **fails**. Free generation beats matched random reach for Llama ($+0.066 [+0.006, +0.142]$) and reverses for gpt-4.1-nano ($-0.061 [-0.129, +0.007]$); the registered condition required both families. The preregistered ablation then removes the remainder: every generated policy lies outside the menu’s support (rate 1.000), and projecting each onto its nearest menu cell makes it indistinguishable from the model’s own menu choice ($-0.000 [-0.042, +0.090]$, $-0.002 [-0.034, +0.025]$). Under the rule we registered, the asymmetry is attributed to output space and the claim is *withdrawn*. What the generator contributes is reach and validity—100.0% of replies parse into an executable policy and all of them run, against 89–91% for matched random—not better search. The second half of C3 passes cleanly in both families: menu selection is significantly worse than matched behaviour-uniform selection from the same menu ($-0.057 [-0.152, -0.021]$, $-0.087 [-0.179, -0.050]$). Free-form generation in full: There is nothing for it to buy. Free-form proposals almost never collapse: across the chains we ran, 0.0 duplicate behaviours per ten proposals under the text-only interface and 0.0 under the execution-grounded one, against a curated menu in which nineteen of forty-one options are one algorithm. Coverage is at its ceiling either way (8.2 unique behaviours per ten proposals with execution feedback against 9.6 without, $p = 0.99$), and the execution-grounded interface produces slightly more configurations that fail to run (1.6 against 0.4). We report this as a negative result with a positive corollary: the redundancy this paper is about is a property of *curated menus*, which is what benchmarks hand to agents, and not of free generation.

P Instrument validation

Three integrity repairs, all registered before their outcome and two of which *failed*. Full artefacts: round5/RESULTS_ROUND5.md, round5/BCA_REPAIR.md.

The mandatory sampled-versus-single-call comparison (R22). Our own amendment replaced sampled elicitation with `top_logprobs` and, in the same document, registered a validation arm as mandatory: on at least 8 landscapes, compare the single-call distribution against at least 40 sampled draws from the same cell, and abandon the endpoint for a family that disagrees beyond the replicate floor. The confirmatory round ran four *other* instrument checks—captured mass, byte-identical repeatability, executable identity of every arm, non-reuse of the old cache—and none of them is this one. It is run here, on the canon arm (the only arm whose menu is exactly regenerable: the rep arms predate a seeding repair and cannot be reconstructed), across all three prompt groups, with the floor computed exactly as the 97.5th percentile of $TV(p, \hat{q}^*)$ over 10,000 draws of $n = 40$ from p . A family passes iff at most 1 of 8 cells disagree.

- **Qwen2.5-1.5B, full-vocabulary logits: PASSES**, 0/8 cells disagreeing, mean excess -0.099 $[-0.121, -0.072]$ —not merely inside the floor but comfortably inside it. The entry-level comparison agrees (0/8).
- **gpt-4.1-nano, top-20 renormalised: FAILS**, 4/8 ($p = 0.00003$), mean excess $+0.025$ $[-0.006, +0.103]$.
- A second registered check—whether the endpoint temperature-scales the alternatives it returns, which the comparison assumes it does not—FAILS for the hosted family: median TV between the $T = 0$ and $T = 1$ reads is 0.090 against a floor of 0.080.
- 0 parse failures in 640 sampled calls.

The registered consequence, fixed before the outcome, is applied: hosted quantities are reported as *not validated by their registered arm* and are demoted, not deleted, and no claim in this paper rests on the hosted family alone. No endpoint, temperature, landscape set, arm or statistic was substituted after the result was seen.

Bounding what a truncated endpoint never showed (R23). A token absent from the returned top-20 has probability no greater than the smallest alternative that *was* returned, so the unobserved permitted mass of a cell is bounded by $U \leq \min((n - |S|)\hat{\epsilon}, 1 - \text{captured})$, and placing all of U on a single behaviour class is more adversarial than any per-identifier allocation. Under the best case U goes to the highest-utility class, under the worst to the lowest. On the 504 local cells U is exactly zero. On the truncated families the largest U anywhere is 0.079 (llama-3.1-8b); the displacement it can produce in a total-variation distance has median 0.00007 and maximum 0.079 over 668 cells. **No headline reverses:** credit stays negative in 0 of 7 models under the best case, the leaderboard stays non-unique on 12/12, and the pairwise flip rate moves from 70.3% to 69.9%.

Executing the order-only attack (R24). PROSPECTIVE. Two targets—one from the family R22 validates and one from the family it fails—eight landscapes, four arms per cell (`ord_best`, `ord_worst`, `written` and a byte-identical replicate), all four elements of the nuisance group acting on the canonical quotient menu and asserted equal as executable sets before any call. The orderings come from the rearrangement inequality applied to the mean text-blind positional profile of the *six other* models, so the adversary has no access to the target; the orderings, the identifier assignment, the predicted utilities and the predicted direction were written to a file and hashed before the first call. **H1 FAILS:** the direction is realised in 10 of 16 cells against a sealed bar of 12 ($p = 0.227$). **H2:** the realised swing is $+0.248$ of the computed ceiling at the median, mean $+0.163$ $[-0.082, +0.474]$ —so the separability assumption partially transports and the R21 ceiling is an overestimate. **H3:** order alone flips the sign of realised credit in 11 of 16 cells. **H4:** the median realised swing $+0.041$ exceeds the median byte-identical floor 0.000, so H1’s failure is a failure of direction rather than noise. Per target, POST HOC: 7 of 8 on Qwen2.5-1.5B (median ratio $+0.385$) against 3 of 8 on gpt-4.1-nano (-0.096).

Is the selector verdict a normalisation artefact? (R25) RE-ANALYSIS, sealed before any alternative was computed. §3.3’s headline is scored on min–max rescaled utilities; that choice was made for readability and never registered as the estimand, and a monotone rescaling is exactly the

kind of unexamined presentational choice this paper is about. The leave-one-landscape-out verdict was recomputed under 4 normalisations of the *same* executed utilities—as published, raw, within-landscape rank, and z —each applied identically to the model arm and the matched-uniform arm, with the min–max path asserted in code to reproduce the published estimator exactly on every cell. **The headline HOLDS:** models beating matched uniform out of sample are 0/0/0/0 of 7 under the four normalisations, and $\rho(p, u)$ is negative in 6/6/6/6. Rank and min–max agree on the sign count (yes), so the result is not an artefact of how far apart the utilities are. The registered response to a reversal was to narrow the claim rather than re-choose the normalisation; none occurred, and the manuscript now states which normalisation its AUC uses, because until this round it did not.

An estimator that was registered but not computed. The sealed analysis plan specifies BCa cluster-bootstrap intervals at 4,000 resamples. Until we audited it, the interval function returned ordinary percentile endpoints, and every analysis importing it inherited the deviation. It is repaired with the bootstrap *draw* unchanged—same generator, same stream, same replicates—so the comparison isolates exactly what the deviation was worth: over 518 intervals, 380 support a jackknife acceleration and are reported as BCa, at a median width ratio of $1.024\times$ and a maximum endpoint movement of 0.130. 138 do not—overwhelmingly the three-prompt-group clustering, where a three-point jackknife is not trustworthy—and those fall back to percentile endpoints, are flagged as such in the artefacts, and are reported as sensitivity analyses rather than as conventional 95% intervals. **10 of 380 intervals change their excludes-zero verdict and no registered verdict changes at all.** The one substantive consequence is stated where it belongs, in §3.3: R16’s no-op-removal control no longer supports the utility relation at conventional significance, though it still supports the agreement and the transfer.

Q Corrections and superseded analyses

- “Free-form generation beats matched random on 12/12 landscapes.” **Retracted.** It measured transport (Appendix O).
- “The landscape is the unit of replication.” **Corrected** for choice-side quantities: there are three independent prompt conditions, not twelve (§2).
- “Quotient normalisation makes the agent a better chooser.” **Not supported, and in one case reversed** (§3.4).
- “C2 passes.” **Retracted.** Computed on five models; it fails on seven (§3.4).
- “The crossing rule predicts the sign in both directions.” **Refuted out of sample**, and the explanation offered for the asymmetry was itself refuted and withdrawn (§4).
- “Three independently authored design spaces.” **Corrected:** grammars B and C share 82–86% of their entries with ours and 86–87% of the coordinate-value vocabulary. **Independence of process does not produce independence of content** (§5).
- “Careful curation shrinks the aliasing surface and enlarges default-dependence, so a curator cannot optimise both.” **Retracted — CONFOUNDED:** the aliasing surface *is* the largest-class size, one quantity under two names.
- “A curator can screen a menu for primacy exposure and jargon density before deploying it.” **Withdrawn.** Registered as R17b and tested: the screen predicts an arrangement’s discovery no better than chance ($-0.131 [-0.385, +0.155]$ over 20 cells and 10 landscapes) and does not beat the one-line baseline of asking where the best option sits ($-0.162 [-0.438, +0.078]$). R16b’s descriptive decomposition stands; the tool built on it does not (§3.3).
- The same test was first registered as R17 on the confirmatory round’s nine arrangements and is **not evaluable** there, for a reason that is a defect in our own instrument rather than a property of the data: those arrangements were never persisted and predate the `stable_seed` fix, so option positions cannot be recovered. R17 scored zero cells, so nothing about the answer was seen before R17b was registered. Any future analysis needing per-arm option positions must re-elicite.
- “An ordering transferred from other models realises its predicted direction when executed.” **Withdrawn.** Registered as R24 with the orderings and the predicted direction hashed before the first call: the direction is realised in 10 of 16 cells against a sealed bar of 12 ($p = 0.227$), and the realised swing is $+0.248$ of the computed ceiling. R21’s ceiling is reported as an overestimate. What survives execution is that order alone flips the sign of realised credit in 11 of 16 cells (§3.3).

- 721 • “The confirmatory round reads the whole choice distribution exactly.” **Corrected.** True of the
722 3 local rungs, whose full identifier logits are read without truncation and which PASSES the
723 registered sampled-versus-single-call validation arm; false of the 4 hosted models, which return
724 at most the top 20 alternatives, renormalised, and which FAILS that arm. Hosted quantities are
725 demoted rather than deleted, and the mass never shown is bounded adversarially without reversing
726 a headline (Appendix P).
- 727 • “The registered intervals are BCa.” **Corrected.** They were percentile intervals until we audited the
728 code against its own protocol. Repaired, published side by side, 10 of 380 intervals change their
729 excludes-zero verdict, and no registered verdict changes (Appendix P).
- 730 • “Removing the no-op class leaves every part of the attractiveness result intact.” **Qualified** by the
731 interval repair above: the agreement (+0.335 [+0.293, +0.377]) and the transfer (7/7) survive
732 that control, the relation to utility (−0.166 [−0.315, +0.002]) no longer excludes zero, and the
733 relation is now claimed on the full menu only (§3.3).
- 734 • “Blind renaming is the dominant channel.” **Qualified.** For three of five models blind renaming
735 raises choice entropy towards uniform, so part of it is information removal rather than rearrange-
736 ment; the pure-nuisance effect of relabelling is measured separately in §4.

737 R Extended limitations

738 **Scope.** The claim is about executable, menu-structured design spaces for closed-loop discovery
739 policies, tested on protein landscapes and tabular AutoML pipelines; free-form program generation
740 of the FunSearch or AlphaEvolve kind is outside what a selection experiment can adjudicate. **Equiv-**
741 **alence is relative to a probe suite.** Policies agreeing on our probes could differ elsewhere; we report
742 held-out-probe splitting (zero), order-sensitive splitting (zero), within-class spread on disjoint seeds
743 (0.0000) and an ϵ sweep rather than asserting the partition is right. **Coverage is uneven.** An API
744 budget was exhausted mid-study; a reduced arm set was used where the full sweep was unaffordable,
745 coverage is in Appendix T, and which cells completed was decided by cost per call, never by result.
746 A concurrency defect in our own reply cache destroyed 26 completed arms, which appear as missing
747 coverage rather than partial data. A sequencing error after a budget failure issued the rewritten arms
748 together with the canonical arm for one family; those cells are labelled retrospective in the sealed
749 file and excluded from the confirmatory test. **The measurement is closing off.** No Llama endpoint
750 above 8B on the router returns per-token alternatives and gpt-5-mini answers a logprobs request
751 with “reasoning is mandatory for this endpoint”; the confirmatory ladder is therefore capped at 8B
752 open-weight rungs plus three OpenAI models. **Two of our own instrument’s registered checks**
753 **fail, and we did not drop them.** The hosted single-call endpoint FAILS the mandatory sampled-
754 versus-single-call validation arm (4/8 cells, $p = 0.00003$) and FAILS the check that it does not
755 temperature-scale what it returns; the local full-vocabulary family PASSES both. Hosted choice-side
756 quantities are therefore instrument-limited, and every headline either holds in the validated family
757 alone or is marked (Appendix P). **Three prompt groups do not support a conventional interval.**
758 Choice-side quantities have three independent prompt conditions, so their cluster intervals cannot
759 carry a trustworthy jackknife acceleration; 138 of 518 intervals fall back to percentile endpoints and
760 are reported as sensitivity analyses. The frequentist precision of the choice-side numbers should be
761 read accordingly. **Ordering coverage remains asymmetric across grammars.** The protein R24
762 intervention executes 16 cells on two models and one prospective ordering. The separately sealed
763 knapsack transport characterises 24 orderings over 6 family prompts and two local Qwen scales, but
764 it does not add a second model family; the grammar was constructed by the researchers. **The repair**
765 **is a measurement contract, not a better searcher.** We show no fixed-budget discovery benefit, and
766 did not look for one after the fact.

767 S Supporting claims and evidential status

768 Table 10 lists supporting claims beyond the four main results in Table 1, using the same evidential-
769 status field and the same generator and canonical files. Registered hypotheses that failed appear here
770 as failures rather than being dropped.

Table 10: Every remaining claim in this work, with its evidential status: the supporting results referenced from the main text, including the registered hypotheses that failed, and the two boundary results of §4. Status labels carry the same three strengths as Table 1: *prospective* means the hypothesis, endpoint and threshold were sealed before the relevant calls existed; *re-analysis* means they were sealed before the quantity was computed, but after the 1,176-cell dataset existed and after related patterns had been seen; *post hoc* means chosen after inspecting the pattern, with descriptive p -values; *audit* is a structural fact about the inputs. Same generator and same canonical files as Table 1.

claim	value	status
distinct prompts sent across the twelve landscapes	3	audit
replicate floor, byte-identical prompts to different landscapes	+0.173 [+0.137, +0.226]	audit
the declared menu’s redundancy as an accidental safety net	gpt-4o-mini: BQN minus raw = -0.503 [-0.683, -0.278]	post hoc
R2-H1 sensitivity vs log parameters, Qwen2.5 ladder	ρ +0.50 CI [+0.50, +0.50]	prospective
registered Gate 4 (BQN $\geq$ behaviour-uniform)	FAIL	prospective
R1-H2 ensemble minus mean single representation (all 7 models)	+0.031 to +0.190; CI above zero in 7/7	prospective
R1-H3 ensemble minus canonical arrangement	FAIL	prospective
landscapes where the top-ranked model is not unique across equivalent menus	12/12	post hoc
pairwise model verdicts that flip sign across equivalent menus	70.3%	post hoc
C3 half 1: free generation vs random at matched reach	FAIL	prospective
C3 half 2: menu selection vs behaviour-uniform	PASS (both families)	prospective
competence vs entropy across all measured models	ρ +0.71, p 0.091	post hoc
competence vs sensitivity across all measured models	ρ +0.46, p 0.31	post hoc
independent replay of the v3 headline numbers	PASS	integrity
R7-H2 the crossing: k-slope of discovery AUC under no-op duplication	llama-3.1-8b -0.033 [-0.079, -0.009] vs gpt-4o-mini +0.178 [+0.101, +0.243]	prospective
R7-H3 parameter-free mixture law	FAIL	prospective
leave-one-model-out robustness of the round’s pooled verdicts	9/11 unchanged under any single-model deletion	re-analysis
R8 crossing rule, cell-by-cell sign accuracy	0.917 (24 cells, $p=1.8e-05$)	post hoc
R9 crossing rule, OUT-OF-SAMPLE sign accuracy on a held-out model	0.750 (12 cells, $p=0.073$)	prospective
R10: does duplication improve the off-default distribution?	FAIL — positive in 2/3 models; llama-3.1-8b slope -0.037 [-0.072, -0.014]	re-analysis
R15: does tempering rescue the selector?	H1 5/7 recover; H2 0/7 beat uniform agreement ρ +0.321; transfers 7/7 held-out models	re-analysis
R16: shared option-level attractiveness	menu position -0.671, jargon +0.603 7/7 models	re-analysis
R16b: what attractiveness is made of		post hoc
R3-H1 task conditioning: removing the task moves the choice above floor		prospective
R3-H3 vs R3-H2: rearranging equivalent options beats retargeting the campaign	$3\times$ to $294\times$, largest on gpt-4o-mini (blind rename TV 0.942 vs wrong task 0.003)	prospective
R17b the curator’s screen predicts an arrangement’s discovery	FAIL — ρ -0.131 [-0.385, +0.155], baseline -0.162	re-analysis
R17 the same test on the confirmatory round’s nine arrangements	NOT EVALUABLE	not evaluable
R18 free pruning: delete the coordinates execution proves inert	8 of 21 coordinates; menu 40 -> 26 options (35%), redundancy 1.87x -> 1.22x	audit

claim	value	status
R19 cost of reporting a single arrangement	AUC error 0.113 (27% of the wording-induced spread); verdict opposite to the all-arm mean 11.7% of the time	post hoc
R20-H3 does doubling the menu steepen the positional slope?	FAIL — paired difference -0.011 [-0.091, +0.033]	re-analysis
R20 blind-renaming control: which peak is position, which is text	first octile 2.85x -> 3.14x; mid-menu peak 2.01x -> 0.54x	post hoc
R21 order-only ceiling: can reordering alone flip the verdict?	84/84 cells straddle the matched null; swing 0.52 of the utility span (best 0.84, as written 0.61, uniform 0.61, worst 0.31)	re-analysis
R21-H2 the order-only ordering transfers without access to the target model	7/7 models; median fraction of the held-out model's own range attained 0.61-0.96	re-analysis
R22 the mandatory A1.2 arm: does a single-call read match ordinary sampling?	local (full-vocabulary logits) PASS 0/8 disagreeing; hosted (top-20, renormalised) FAIL 4/8 (p=0.00003)	prospective
R22-H3 does the hosted endpoint temperature-scale the alternatives it returns?	FAIL: median TV 0.090 vs a replicate floor of 0.080	prospective
R23 adversarial bound on the mass a truncated hosted endpoint never showed	max unobserved permitted mass 0.079; TV displacement median 0.00007, max 0.079; headlines reversed: 0 of 7	re-analysis
R24 the order-only attack EXECUTED: is the predicted direction realised?	FAIL: 10/16 cells against a sealed bar of 12 (p=0.227)	prospective
R24-H2 how much of R21's computed ceiling an executed reordering realises	median realised/computed swing +0.248; mean realised swing +0.163 [-0.082, +0.474]	prospective
R24-H3 does reordering alone flip the sign of REALISED credit?	11/16 cells	prospective
the registered BCa cluster intervals, computed at last	380 of 518 intervals support BCa; 10 change their excludes-zero verdict; 0 registered verdicts change	integrity

771 T Per-cell results and coverage

772 Table 15 lists every model×landscape cell with a completed transformation sweep, with its median
773 primary-family distance, its same-prompt replicate floor, its canonical credit, the range of credit over
774 all rewritings, and the number of rewritings that reverse the verdict. Cells without a completed sweep
775 are absent rather than partially reported.

776 U Exploratory robustness and transport analyses

777 These analyses were conducted after the confirmatory record was frozen. They remain separate from
778 that record and do not alter the four headline claims. We report only additions that perform a distinct
779 scientific function; redundant plotting-only re-expressions and closeout counts are omitted.

780 U.1 Generated rewrite styles

781 One Qwen2.5-0.5B generator rewrote each sealed option under three instructions (terse, tutorial, and
782 jargon-preserving), using separate sampling seeds; Qwen2.5-1.5B then scored the fixed identifiers
783 and behavioural classes. These are three generated rewrite styles, not independent human authors.

784 U.2 Cost–stability and heterogeneity

785 Two descriptive views of the confirmatory record ask how much repeated arrangement sampling
786 buys and whether representation sensitivity is uniform across model families. Figure 11 reports the
787 registered cost–stability points; Figure 12 reports model-by-transformation heterogeneity.

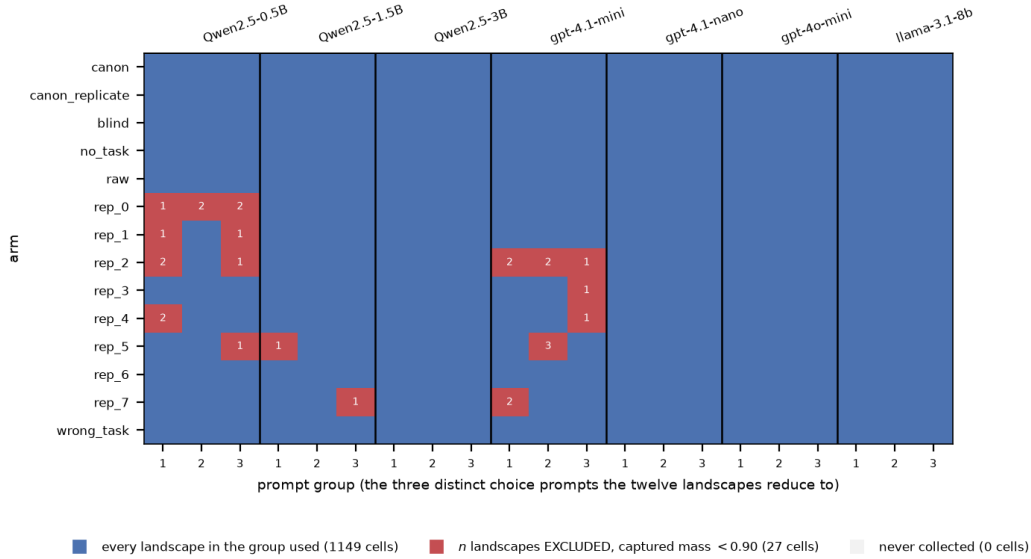


Figure 6: **Every cell of the confirmatory round, and its disposition.** Rows are arms, columns are model $\times$ prompt group—the prompt group rather than the landscape, because this round’s own audit showed the 12 landscapes send only 3 distinct prompts. 0 cells were never collected. 27 of 1176 are present but fall below the registered 0.90 captured-mass floor and are therefore disqualified from every analysis; the number in a red cell is how many landscapes of that group are excluded. **every one of them a representation arm**, spread over 7 of the eight representation arms: *no* cell of canon, canon_replicate, blind, raw, no_task or wrong_task—the arms every headline contrast is measured on—is missing or excluded.

Table 11: **Logit-round evidence and population.** Table 1 states what each result claims and how strongly it is registered; this counts the logit-round cells; sampled-round populations are in the main table. Counts are read from the artefacts by the generator, and apply the registered captured-mass floor, so a population cannot drift from the analysis that used it. The final row covers the two boundary results of §4.

	result	population	what it does not cover	status
R1	The written menu is not the executable space	40–41 executable policies per landscape from a 41-policy declared pool; 21 declared coordinates	no model in the loop — execution only	audit
R2	Equivalent rewritings change credit and which model appears best	252 cells: 7 models $\times$ 12 landscapes $\times$ 3 execution-verified arms	27 cells excluded by the captured-mass floor, all of them rep_* arms ; 0 never collected	prospective
R3	The models read the page, and their ordering is harmful	position: 84 cells on the two arms whose identifiers encode position; tempering: 84 canon cells; executed reordering: 16 sealed held-out cells on 2 targets	the executed reordering covers 2 of 7 models by design and by budget	re-analysis; the executed arm prospective
R4	Stable measurement does not guarantee better selection	C2: 84 quotient-menu cells against their raw-menu twins; ensembling: 645 cells over 8 arrangements	the ensembling arms carry every excluded cell in the round	prospective
B	Boundaries: task conditioning, and the redundancy crossing	task controls: 168 cells, 7 models; dose-response: 2 models chosen before any dose call	two models cannot separate <i>the rule works</i> from <i>these two models differ</i>	prospective; the sign rule descriptive

Writing one behaviour more ways changes what is chosen; the executable set never changes

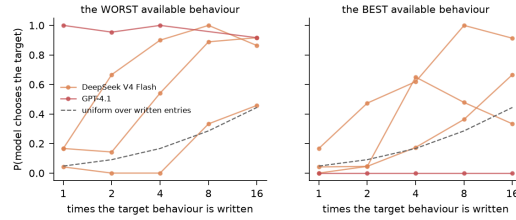


Figure 7: Dose-response, with the executable set verified fixed entry by entry. Dashed: what uniform selection over written *entries* would give. A menu read as a set of behaviours would be flat.



Figure 8: Asked which of the declared options execute identically, models recover little of the true quotient. Bars are pairwise recall against the executed partition; a correct audit reaches 1.0. Missing bars are cells where the API budget stopped the run.

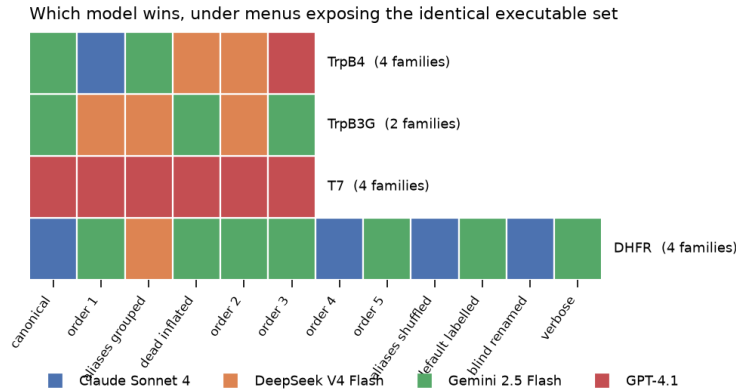


Figure 9: Which model an evaluation would name best, per landscape, under each written space. Every column exposes the identical executable opportunity set.

788 U.3 Exploratory transport to the existing AutoML grammar

789 Figure 13 applies the same representation interventions to three pre-existing AutoML menus. This
790 checks mechanism transport only; it is not a new domain-level performance claim.

791 **Repair remained unsuccessful.** The separately sealed repair test did not clear its promotion gate:
792 arrangement ensembling reduced the credit-sign flip rate by only 20% relative to a single written
793 menu, below the declared 50% threshold, and the utility/direction requirements were not jointly
794 satisfied. Behavioural quotienting therefore remains a measurement contract, not a demonstrated
795 improvement to discovery.

Table 12: The sealed prediction, scored on cells whose rewritten menus did not exist when the rule was registered. The rule and its coefficients were fitted on the development slice alone.

prediction	result	verdict
P1 high- vs low-risk reversal rate	26% vs 15%, ratio 1.7, $p = 0.227$	FAIL
P2 margin ranks cells	$\rho = +0.41$, $p = 0.103$, $n = 11$	FAIL
P3 calibration	Brier 0.157 vs base rate 0.159	PASS
P4 margin beats landscape crowding	AUC 0.64 vs 0.54	PASS

Table 13: Representation Invariance Score. RIS = 1 means the median tested distance does not exceed its byte-identical replicate floor; it does not certify invariance. The statistic summarizes three primary rewrites and available additional orderings, with at least three eligible rewrites per cell. Upper-tail distances and unequal model coverage are reported in Appendix V. Invariance and utility are different quantities.

model family	landscapes	mean RIS	mean RIS-C (credit)
GPT-4.1	3	0.944	0.974
DeepSeek V4 Flash	12	0.703	0.936
Claude Sonnet 4	3	0.690	0.861
Gemini 2.5 Flash	10	0.368	0.910

Table 14: Sampled-round arm coverage. Each entry counts arms with at least eight valid class replies, including canonical and replicate arms. Missing sweeps are not counted as complete.

landscape	Claude Sonnet 4	DeepSeek V4	Gemini 2.5	GPT-4.1
DHFR	13	13	13	13
GB1	2	10	13	2
ParD2	2	11	13	2
ParD3	2	7	13	2
T7	13	7	13	13
TEV	2	7	13	2
TrpB3D	2	11	13	2
TrpB3E	2	7	13	2
TrpB3G	0	7	13	0
TrpB3H	0	7	4	0
TrpB3I	0	7	0	0
TrpB4	13	7	13	13

796 U.4 Registered order-population transport to a knapsack grammar

797 To test whether the order result was a favourable single permutation, we constructed a distinct
798 executable grammar of eight deterministic knapsack heuristics and six standard instance families.
799 Exact execution on 64 probe instances per family verifies eight distinct canonical behavioural classes;
800 every written arm has the same quotient. Held-out utility is mean value divided by the dynamic-
801 programming optimum on 128 disjoint instances per family and never enters a prompt or signature.
802 Before model scoring, we sealed two cached local models, 24 independently seeded line orders,
803 six outcomes and six gates. The six families are the natural units; permutations are intervention
804 replicates.

Table 15: Every model $\times$ landscape cell with a completed transformation sweep. **med. TV**: median total-variation distance from the canonical menu over the primary family; **floor**: the same distance between two independent batches of the *byte-identical* canonical prompt; **C**: credit against the behaviour-uniform null on the canonical menu, and its range over all rewritings; **rev.**: rewritings that reverse the sign of *C*.

model	landscape	med. TV	floor	<i>C</i> (canonical)	<i>C</i> range over rewritings	rev.
Claude Sonnet 4	DHFR	0.792	0.042	+0.126	[−0.649, +0.134]	5/11
Claude Sonnet 4	TrpB4	0.083	0.250	−0.103	[−0.571, −0.036]	0/11
Claude Sonnet 4	T7	0.417	0.083	+0.900	[+0.322, +0.989]	0/11
DeepSeek V4 Flash	DHFR	0.375	0.125	−0.032	[−0.413, +0.115]	4/11
DeepSeek V4 Flash	TrpB4	0.375	0.208	−0.107	[−0.464, +0.043]	2/5
DeepSeek V4 Flash	T7	0.458	0.208	+0.540	[+0.005, +0.761]	0/5
DeepSeek V4 Flash	GB1	0.375	0.208	−0.575	[−1.244, −0.059]	0/8
DeepSeek V4 Flash	ParD2	0.333	0.236	−0.001	[−0.001, +0.001]	4/9
DeepSeek V4 Flash	ParD3	0.542	0.292	−0.001	[−0.005, +0.001]	1/5
DeepSeek V4 Flash	TEV	0.333	0.167	−0.053	[−0.064, −0.002]	0/5
DeepSeek V4 Flash	TrpB3D	0.375	0.167	−0.145	[−0.381, +0.013]	1/9
DeepSeek V4 Flash	TrpB3E	0.500	0.250	+0.235	[−0.060, +0.489]	2/5
DeepSeek V4 Flash	TrpB3G	0.458	0.167	−0.128	[−0.464, +0.080]	2/5
DeepSeek V4 Flash	TrpB3H	0.500	0.208	−0.113	[−0.151, +0.091]	1/5
DeepSeek V4 Flash	TrpB3I	0.292	0.208	−0.192	[−0.523, −0.048]	0/5
Gemini 2.5 Flash	DHFR	0.708	0.167	−0.077	[−0.403, +0.117]	8/11
Gemini 2.5 Flash	TrpB4	0.583	0.083	−0.024	[−0.376, +0.094]	2/11
Gemini 2.5 Flash	T7	0.583	0.208	+0.792	[−0.027, +0.862]	1/11
Gemini 2.5 Flash	GB1	0.333	0.125	−0.455	[−1.934, +0.245]	1/11
Gemini 2.5 Flash	ParD2	0.875	0.167	−0.000	[−0.003, +0.000]	1/11
Gemini 2.5 Flash	ParD3	0.708	0.083	−0.004	[−0.005, +0.001]	4/11
Gemini 2.5 Flash	TEV	0.292	0.083	−0.033	[−0.064, +0.023]	1/11
Gemini 2.5 Flash	TrpB3D	0.833	0.000	−0.126	[−0.445, +0.086]	3/11
Gemini 2.5 Flash	TrpB3E	0.708	0.292	+0.072	[−0.079, +0.643]	4/11
Gemini 2.5 Flash	TrpB3G	0.792	0.083	−0.005	[−0.464, +0.219]	8/11
Gemini 2.5 Flash	TrpB3H	0.750	0.208	−0.149	[−0.149, −0.019]	0/2
GPT-4.1	DHFR	0.000	0.000	−0.649	[−0.649, −0.042]	0/11
GPT-4.1	TrpB4	0.000	0.000	−0.571	[−0.571, +0.066]	1/11
GPT-4.1	T7	0.083	0.000	+0.989	[+0.749, +0.989]	0/11

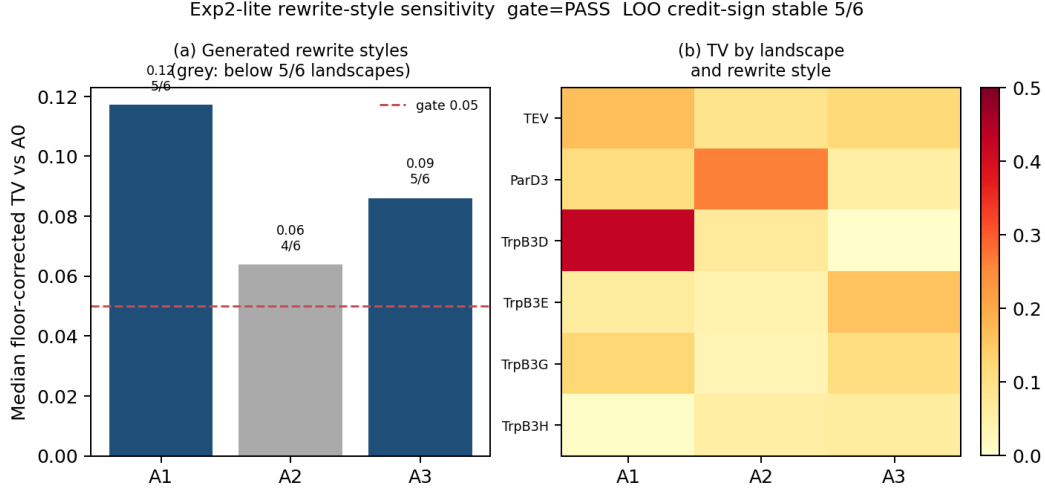


Figure 10: **Sensitivity to generated rewrite style.** The terse and jargon-preserving styles exceed the declared 0.05 TV threshold on 5/6 landscapes; the tutorial style does so on 4/6. The top class changes once across 18 style–landscape cells, and the credit sign changes in three cells, all on the same near-zero-credit landscape. The result supports wording sensitivity under one local generator/scorer pair, not population-level author generality.

Action path: more arrangements stabilize credit; uniform still competitive

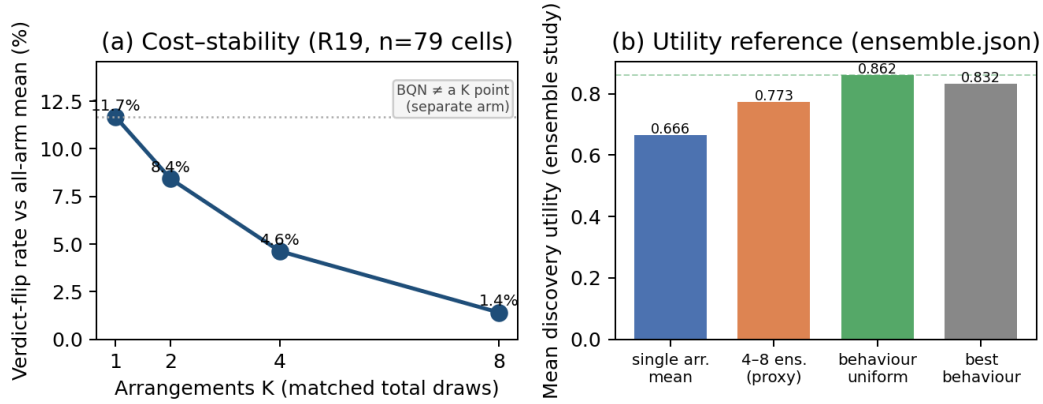


Figure 11: **Cost-stability frontier from the registered arrangement experiment.** Verdict disagreement with the all-arrangement mean falls from $K=1$ to $K=8$ over 79 cells. The behavioural quotient is shown separately because it is not another arrangement count. This is an actionability display of existing evidence; it does not establish an optimum outside the tested values or a discovery-utility gain.

H1: floor-aware transform sensitivity (median over landscapes)
* hosted top-20 = secondary

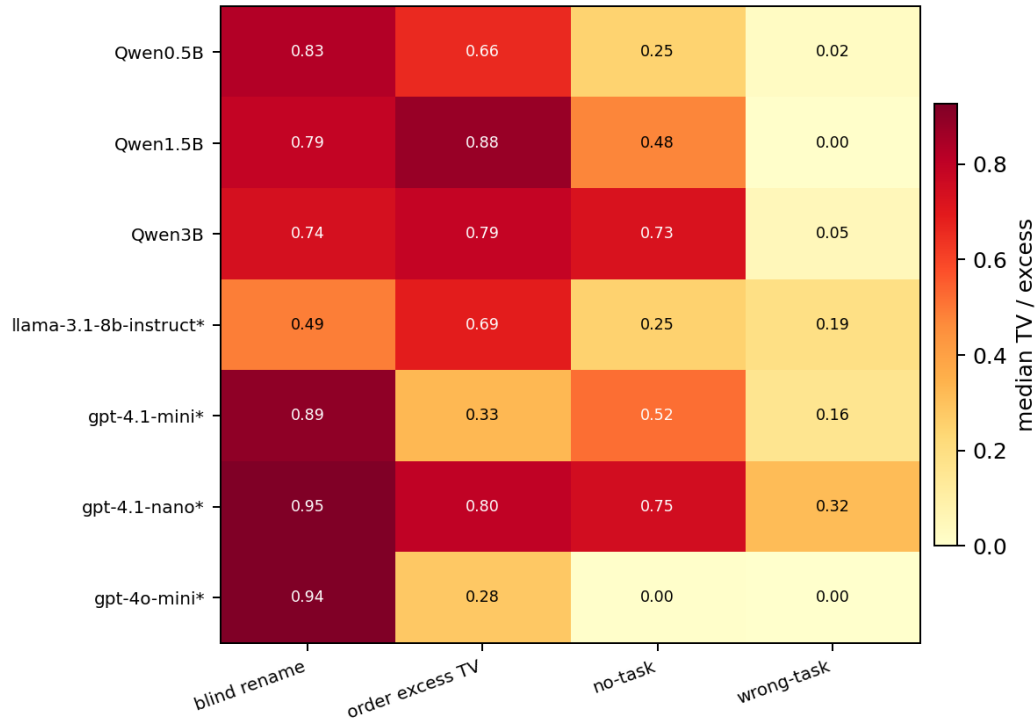


Figure 12: **Model-by-transformation heterogeneity in the confirmatory record.** Cells are median floor-aware sensitivities across landscapes. Local full-vocabulary-logit models are primary; starred hosted rows are secondary because their top-20 instrumentation failed a registered validation check. This is a descriptive re-expression of the confirmatory data, not an additive variance decomposition.

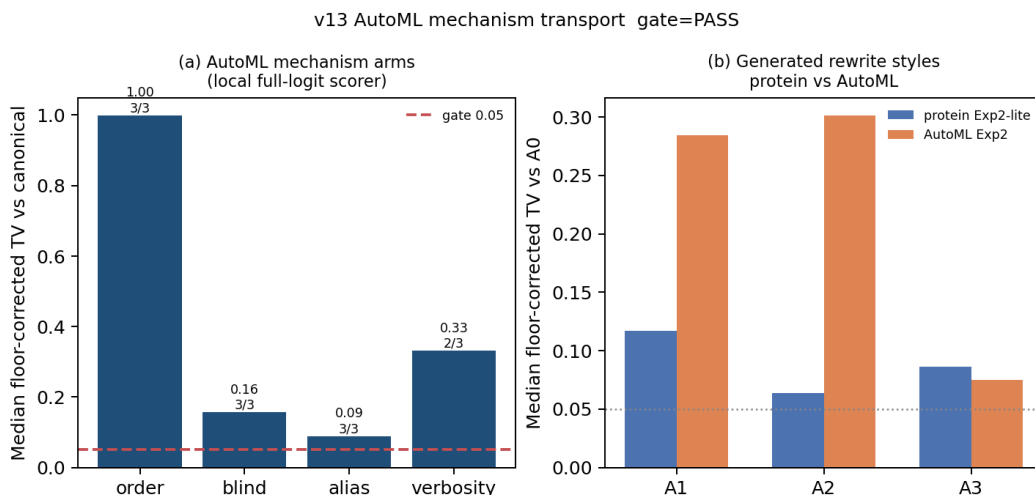


Figure 13: **Representation transforms on three existing AutoML menus.** Under one local full-logit scorer, order permutation produces near-maximal TV and top-class changes on all three datasets; blind renaming changes the distribution without changing the top class. The second panel compares the same three generated rewrite styles across protein and AutoML menus. With three datasets and one scorer, this is exploratory mechanism transport, not a new independent-domain headline.

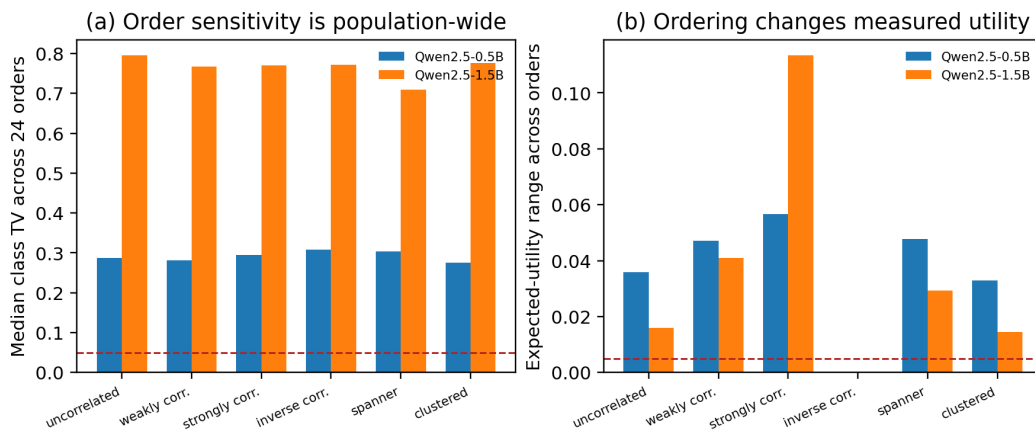


Figure 14: Population-level order transport in a third executable grammar. (a) For Qwen2.5-0.5B/1.5B, the median across family-level median TV distances is 0.291/0.771; all 6/6 family medians exceed the registered 0.05 threshold and the smallest leave-one-family-out medians are 0.287/0.770. (b) Ordering changes expected held-out utility by more than 0.005 on 5/6 families for each model; the resulting credit interval crosses the behaviour-uniform null on 6/6 families for Qwen2.5-0.5B and 1/6 for Qwen2.5-1.5B. All 6/6 gates pass. The grammar was constructed by us, so this is preregistered domain/grammar transport rather than independent authorship.

V Corrected population accounting and RIS interpretation

The sampled round contains 8,088 retained calls, including 31 invalid class responses. Canonical data cover 41 of 48 possible four-model by twelve-landscape cells. The primary-family analysis covers 29 cells: Claude Sonnet 4 contributes 2/9 credit reversals, DeepSeek V4 Flash 4/36, Gemini 2.5 Flash 6/31, and GPT-4.1 0/9, totalling 12/85. Four additional orderings and shuffled aliases raise the primary-family total to 38/193. Secondary rewrites contribute 18/62; 56/255 is the combined count. These are empirical sign changes, not tests that both credit estimates separately exclude zero. Invalid replies are excluded from the conditional class distribution; credit requires at least eight valid replies in each compared arm. Different arm coverage prevents interpreting pooled rates relative to the 3/41 replicate rate as a matched causal risk ratio.

RIS uses the three primary arms plus available additional orderings, requiring at least three eligible rewrites and a same-prompt replicate. The 28 eligible cells are distributed unevenly: 3 Claude, 12 DeepSeek, 10 Gemini and 3 GPT-4.1. Their model-mean RIS values are 0.690, 0.703, 0.368 and 0.944; their largest tested TV distances are 1.000, 0.833, 1.000 and 0.917, respectively. One Claude cell and one GPT-4.1 cell have $RIS = 1$ while at least one tested distance exceeds their replicate floor. Thus the loss of a universal invariance guarantee is observable in these data, not just in the $(0, 0, 1)$ counterexample. The statistic and its historical values are unchanged; the corrected interpretation reports the finite transformation population and upper-tail behavior.

W Evidence-informed selection qualification

A separately sealed qualification tests whether prior experimental performance supplies useful information to the selector. It uses the 21 class representatives common to the retained twelve landscapes. For each target, prior scores average normalized policy utilities equally across the other five biological groups; all TrpB landscapes form one group, both ParD landscapes another. No target-group utility is shown. The two cached Qwen2.5-0.5B/1.5B models each receive four fixed order permutations with and without those scores: 192 full-vocabulary conditional identifier distributions, all retained. The 12 landscapes provide utility contexts; informed prompts comprise only six distinct evidence states, or 24 strings across four orders. Without evidence there are four distinct strings per model. This uses already inspected benchmark utilities, so it is a qualification test, not an untouched confirmation or an independently curated menu.

The reference choosing the largest prior mean has normalized credit -0.016 $[-0.102, +0.072]$ and fails the prespecified competence gate. The paired effect of supplying evidence is -0.005 $[-0.016, +0.003]$ for 0.5B and -0.054 $[-0.139, +0.007]$ for 1.5B. Intervals are percentile bootstraps over six biological groups, with 10,000 resamples; they do not quantify uncertainty over models or grammars. Informed selection also fails the model competence gate: credit $+0.007$ $[-0.010, +0.022]$ and -0.012 $[-0.046, +0.022]$, respectively. Order sensitivity persists, but this failed information-transfer reference does not establish that a competent scientific selector loses ability under a rewrite. The following same-landscape qualification was sealed separately before its pilot executions and calls.

One local repair: same-landscape pilot evidence. To test useful information before expanding the model panel, we selected the lexicographically first landscape from each biological group: DHFR, GB1, ParD2, T7, TEV and TrpB3D. The same 21 representatives each run three fresh seeds (700–702): 378 campaigns. Pilot mean fitness is normalised across policies using only pilot outcomes; evaluation utilities on seeds 100–119 stay hidden. The largest-pilot-mean reference has positive credit on four of six landscapes and mean $+0.115$ $[-0.074, +0.369]$, passing the prespecified admission gate (positive mean and at least four positive landscapes). That gate is a feasibility condition, not a significant performance claim. We then collected all 96 planned distributions from the same two models, four orders and paired with/without-evidence conditions.

For Qwen2.5-0.5B, mean credit changes from $+0.009$ $[-0.005, +0.024]$ without evidence to $+0.012$ $[-0.001, +0.025]$ with it. For 1.5B, it changes from $+0.018$ $[-0.035, +0.102]$ to $+0.080$ $[-0.017, +0.193]$. Paired utility changes are $+0.003$ $[-0.004, +0.009]$ and $+0.062$ $[-0.011, +0.111]$. Neither informed model clears the competence criterion (positive lower interval bound). Mean TV to the first ordering is 0.676/0.793 under evidence, and credit crosses zero across orderings on 5/6 and 4/6 landscapes, respectively. These conditional distributions show continued

sensitivity; they do not establish a loss of competent scientific selection. Conditional identifier probabilities use the full local vocabulary; unrestricted permitted-token mass is retained for every call, and no low-mass call is silently dropped. Model-scale comparisons and six-landscape intervals remain limited.

The cross-group and same-landscape tests are separate and are not pooled. Both use already inspected target utility contexts and the original researcher’s class pool; no independent curation or frontier-model validation is claimed. Pilot evidence would cost $21 \times 3 \times 120$ assay observations per landscape in a laboratory, supplied equally to each compared selector here. Computation uses published measured landscapes, not new wet-lab assays. The failed cross-group reference and both non-significant model gains remain reported.

X When default-mass transfer explains duplication

The opposite effects in Table 7 do not establish a predictive law. We test a simple rival: duplication changes the default’s probability while preserving relative probabilities among the other classes. This is a distributional hypothesis, not a new theorem. Historical results below are development analyses of previously inspected responses and fixed utility means.

Separate attraction from redistribution. Let p and q be baseline and duplicated-menu distributions, d the default, and $\mu_{-d}(p)$ the expected utility conditional on a non-default draw. For $p_d, q_d < 1$, direct algebra gives

$$E_q[u] - E_p[u] = (q_d - p_d)\{u_d - \mu_{-d}(p)\} + (1 - q_d)\{\mu_{-d}(q) - \mu_{-d}(p)\}. \quad (2)$$

The first term changes default mass; the second redistributes the remaining mass. Dropping the second term requires a checked condition. The mass-matched diagnostic sets $\hat{q}_d = q_d$ and $\hat{q}_c = (1 - q_d)p_c / (1 - p_d)$ for $c \neq d$. It is the attraction mixture $(1 - a)p + ae_d$ only when $a = (q_d - p_d) / (1 - p_d) \in [0, 1]$. A decrease in default mass violates this condition. It uses the observed intervention mass and therefore cannot predict that mass.

The repeated-selection endpoint is nonlinear. For sorted class utilities and cumulative probabilities F_j , $E[\max_{i \leq n} u(B_i)] = \sum_j u_j (F_j^n - F_{j-1}^n)$. We evaluate this expression at the nine registered draw counts before integration. Interpolating default and off-default AUC values instead draws the mixture component once for an entire sequence; it is not the experiment’s independent redraw at each selection. The error persists even if non-default ratios are perfectly preserved. All AUC errors below use the exact CDF calculation.

A predictor that uses only the baseline. An equal-weight alias rival predicts $\tilde{q}_d = kp_d / [1 + (k - 1)p_d]$ and $\tilde{q}_c = p_c / [1 + (k - 1)p_d]$ for $c \neq d$. It has no fitted parameter or post-intervention input. We compare its AUC error with predicting no change. Predictions condition on previously audited class utilities; they do not predict unknown fitness. This tests the proposition that aliases preserve per-entry choice weights; contextual language-model probabilities need not satisfy it.

Table 16: Distribution-level explanation and prediction of alias effects. TV is between conditional non-default distributions; errors are absolute errors in normalised selection AUC, averaged over all non-baseline doses. Matched mass uses intervention data; odds and no-change use only baseline data.

Panel	Model	N	TV (median)	Matched	Odds	No change
AutoML (old)	Llama-3.1-8B	12	0.236	0.018	0.022	0.015
AutoML (old)	GPT-4o-mini	12	0.096	0.014	0.022	0.022
Protein (old)	Llama-3.1-8B	48	0.414	0.062	0.074	0.080
Protein (old)	GPT-4.1-nano	48	0.842	0.112	0.153	0.153
Protein (old)	GPT-4o-mini	48	0.003	0.003	0.424	0.425
AutoML (fresh)	Qwen2.5-0.5B	24	0.659	0.046	0.061	0.067
AutoML (fresh)	Qwen2.5-1.5B	24	0.587	0.029	0.047	0.036

Historical results and their limit. GPT-4o-mini’s protein responses are consistent with near-preserved non-default ratios, but Llama and nano strongly redistribute that mass. Supplying the observed default mass gives the former a small AUC error; predicting that mass from multiplicity fails. The two statements are compatible: a good conditional explanation does not establish a usable predictor. The AutoML diagnostic also does not uniformly beat no change. These panels inherit the hosted-logit validation limitation. Protein doses share three prompt templates; the 48 contrasts per model are not 48 independent tasks. We report descriptive errors, not confidence intervals based on pseudo-replication. The previous failed sealed prediction and AutoML transfer retain their original status.

Fresh local challenge. Before new inference we froze both predictors, two cached Qwen scales, three existing AutoML datasets (wine, digits, breast cancer), two placements of the default (first/last), and $k \in \{1, 3, 6, 10, 19\}$. Each alias repeats exactly the canonical default text; equivalence follows by dispatching it to the same executed policy. Every task retains its original classes and fixed utilities. The dataset’s row, feature and class counts are supplied; utilities are withheld. Byte-identical repetitions at $k = 1$ and 19 give 84 forward passes in total. All permitted one-token identifiers are scored from the full vocabulary. This measures constrained identifier selection, not unrestricted generated replies. Allowed-token mass ranges from 0.983 to 1.000; maximum repeat TV is 0. Doses and placements are repeated measures within three tasks, and the two scales share a model family. These are new responses, not new datasets, independent grammar curation or demonstrated competent scientific selection. Placement changes the effect of absolute positions, identifiers and context length together; it does not separately identify those causes.

The frozen distributional tolerance requires mean matched TV at most 0.05 in every task/placement cell. The prospective criterion requires mean AUC error at most 0.02 and no worse than no change in every such cell. These are operational tolerances, not significance thresholds or a universal definition of usefulness. For Qwen2.5-0.5B, the distributional criterion fails and the prospective criterion fails. For Qwen2.5-1.5B, the distributional criterion fails and the prospective criterion fails.

Table 17: All fresh task/placement cells; four doses averaged within each cell. No tasks or failures are excluded. Matched TV concerns the full class distribution; the final three columns are mean absolute selection-AUC errors.

Model	Task	Default	Matched TV	Matched	Odds	No change
Qwen2.5-0.5B	breast_cancer	first	0.217	0.039	0.080	0.097
Qwen2.5-0.5B	breast_cancer	last	0.300	0.159	0.138	0.138
Qwen2.5-0.5B	digits	first	0.310	0.044	0.070	0.083
Qwen2.5-0.5B	digits	last	0.245	0.023	0.054	0.054
Qwen2.5-0.5B	wine	first	0.246	0.011	0.012	0.017
Qwen2.5-0.5B	wine	last	0.340	0.002	0.013	0.013
Qwen2.5-1.5B	breast_cancer	first	0.639	0.059	0.108	0.089
Qwen2.5-1.5B	breast_cancer	last	0.742	0.050	0.048	0.050
Qwen2.5-1.5B	digits	first	0.625	0.030	0.021	0.042
Qwen2.5-1.5B	digits	last	0.260	0.012	0.064	0.015
Qwen2.5-1.5B	wine	first	0.509	0.023	0.020	0.022
Qwen2.5-1.5B	wine	last	0.239	0.002	0.017	0.002

The complete prompts, class mappings, model revisions, permitted-token log probabilities, distributions, source hashes and seal are retained. No model or threshold was changed after inference. Negative criteria narrow the explanatory claim; they do not erase the observed causal effects of menu representation.