
Do Repeated LLM Decisions Improve CRISPR Hit Discovery?

Odysseus Kaloidas
UC Berkeley

Abstract

1 Does an LLM that benefits from experimental feedback also choose better ex-
2 periments than an informed numerical policy? We study retrospective CRISPR
3 hit discovery, where a nearest-neighbour procedure expands five proposed genes,
4 called centres, into a batch. Our numerical rival samples these centres from pre-
5 viously observed hits while retaining the same initial batch, features, expansion
6 procedure and test budget. Across six models and six readouts from four sources,
7 correct outcomes improve LLM yield over withheld outcomes by 18.18 hits per
8 640 tests, yet the rival finds 9.98 more hits. Complete-campaign checks with
9 expanded response budgets and a native JSON interface retain this pooled distinc-
10 tion. The native-interface model returns valid proposals without replacement of
11 missing genes, but wins on one source and loses on another. Initialization also
12 changes which controller performs better. Finally, LLM proposals contain some
13 useful candidates that a hypothetical selector with access to future outcomes could
14 exploit, but the tested selectors do not establish a gain over equally sized numeri-
15 cal candidate pools. These conditional findings show why evaluating added LLM
16 yield requires an alternative retaining both outcome information and numerical
17 batch expansion.

18 1 Introduction

19 Does feedback help an LLM choose experiments, and does the LLM use that feedback better than a
20 simple rule? These are different questions. Removing or reassigning outcomes compares the LLM
21 with itself. A numerical rule may use correct outcomes just as effectively by selecting a known hit
22 and testing its untested neighbours. We therefore compare the LLM with that informed alternative,
23 not only an uninformed random baseline.

24 In our retrospective CRISPR task, a test reveals a gene’s outcome from a completed screen. The
25 LLM chooses five genes, called centres; a nearest-neighbour (NN) procedure fills the batch with
26 untested genes having similar feature profiles. We call the centre-selection rule the controller and
27 the expansion procedure the executor. Our numerical rival, *uniform-hit*, samples centres uniformly
28 from previously observed hits and uses the same executor. Both controllers inherit the same LLM-
29 selected first batch and its outcomes: this is their shared initialization. Uniform-hit recovers 73.92
30 hits per 640 tests against the LLM’s 63.94. The comparison therefore asks who should choose
31 subsequent centres, not whether to replace the LLM’s initial gene selection.

32 We separate feedback benefit from controller advantage, examining source and initializer reversals
33 and the gap between useful proposals and successful selection. We test the centre–NN architecture
34 of Gupta et al. [2025] under matched controls, without BioDiscoveryAgent’s search and critic tools.

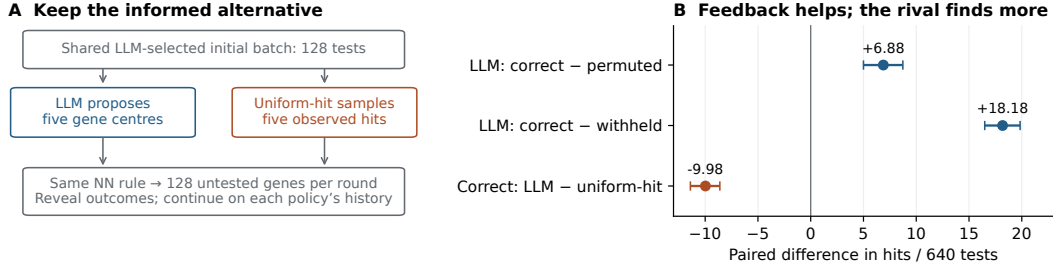


Figure 1: Matched access to observed outcomes, initialization and NN expansion. Policies share 128 initial tests, then develop their own histories. Right: three paired contrasts from 720 starts with 98.333% intervals; magnitude depends on source. Centres guide 128 new tests; tested centres are not retested.

2 Related work

BioDiscoveryAgent combines LLM control with search and critic tools [Roohani et al., 2025]. Its main evaluation also tests LLM initialization followed by Coreset (Section 5, Table 1 of that paper); handoff alone is therefore not our contribution. Gupta et al. [2025] report little sensitivity to randomized feedback in their BioDiscoveryAgent tests, then introduce NN acquisition with complete campaigns and a random-centroid ablation. Random centres remove informed anchor selection along with the LLM; uniform-hit retains it. Wainrib et al. [2026] establish positive feedback effects in Cell Painting discovery. Our comparison retains feedback while replacing the LLM with an informed rival. Different models, output requirements and acquisition procedures make this a complementary test, not a literal replication overturning those studies.

GeneDisco shares source screens with us [Mehrjou et al., 2022]. AssayLoop’s learned-policy hand-off uses different historical training labels [Edwards et al., 2026].

SOBER and LLAMBO separate representation, initialization, prediction and proposal roles [Kristiadi et al., 2024, Liu et al., 2024]. Positive optimization results include LLM operators within evolutionary selection (MoLLEO), preference guidance (LGBO), and selection among numerical acquisition functions (LMABO) [Wang et al., 2025, Yuan et al., 2026, Ngo et al., 2026]. These roles differ from choosing gene centres; our comparison does not establish that continued LLM guidance is generally ineffective.

Self-correction studies separate the effects of feedback, repeated sampling and critique [Huang et al., 2024]. External verification can improve iterative prompting without informative verbal critiques [Stechly et al., 2025]. With measured screen outcomes, our question is whether LLM control improves acquisition over a policy that retains those outcomes and the same executor.

3 Methods

3.1 Task and acquisition procedure

The original panel comprises Carnevale adenosine, Sanchez tau/tau-down, Schmidt IL2/IFNG and the released CAR-T screen [Carnevale et al., 2022, Sanchez et al., 2021, Schmidt et al., 2022, Roohani et al., 2025]. A readout is one screened outcome; a source is its originating study. These six readouts span four studies: Schmidt and Sanchez each supply two, so readouts are not independent sources. A B640 campaign starts with 128 tests, then adds four 128-gene adaptive batches, for 640 distinct genes. Scharenberg choline recycling is a separate study with five batches of 32 genes (B160) [Scharenberg et al., 2023]; screen and budget change together. Achilles features summarize gene-disruption effects across cell lines [BioDiscoveryAgent contributors, n.d., Broad Institute, n.d., Tsherniak et al., 2017, Meyers et al., 2017]. NN treats similar profiles as a proxy for related target-screen responses: a transfer assumption. B640 pools contain 17,009–17,758 eligible genes and 139–883 released hits; Scharenberg has 1,013 genes and 48 hits (Table 13). We retain each source’s released hit labels (its topmovers array), rather than reconstructing a common threshold. Each acquired gene contributes one hit if listed, zero otherwise; no LLM judge scores the expla-

nation. Zero denotes an unlisted gene, not a validated biological negative. Sanchez tau uses the released bidirectional task description; processed score signs are not interpreted as biological effect directions. Appendix A.1 binds actual prompts to evaluated labels. Tests reveal outcomes from already completed screens; these campaigns conduct no new physical perturbations.

Llama3.1-8B, Qwen2.5-7B, Gemma3-4B, Ministral8B-2512, Qwen3-8B and Phi-4-14B provide the original model panel [Grattafiori et al., 2024, Qwen et al., 2024, Team et al., 2025, Mistral AI, 2025, Yang et al., 2025a, Abdin et al., 2024].

Table 1 identifies the population supporting each main comparison. These studies are not a fully crossed eight-model, eight-readout experiment.

Table 1: Population and output requirements for each comparison. Starts pair campaigns, not biological replicates. Except for the varied factor, policies share LLM initialization, correct outcomes, NN expansion and B640 (Scharenberg: B160). Original responses use three fields and 384 tokens; changes below give per-attempt limits.

Comparison	Models / output format	Sources/readouts	Starts
Feedback / control / executor	Original six	Four studies/six readouts	720
Source and budget	Original six	Scharenberg	120
External controller	Original six	ORCS1107	120
External feedback	Qwen2.5/Phi-4	ORCS1107	40
Initializer $\times$ controller	Gemini/DeepSeek; JSON, 4096/2048 tokens	Carnevale/IL2	80/model
Longer list responses	Qwen2.5/Ministral; list, 1536 tokens	Carnevale/IL2	80
Native JSON responses	Gemini; JSON, 4096 tokens	Carnevale/IL2	40
Proposal selection	Llama/Qwen2.5	Original readouts; Scharenberg	120;20

80

81 Within each computational start, policies share the realized LLM-selected initial batch. They then
82 choose their own genes, so matched budgets do not imply identical later histories. Starts repeat
83 computational acquisition, not biological assays. Models have different initial batches, so their
84 comparison does not isolate capacity.

85 The prompt describes the screen, complete acquired history and NN expansion, separating [HITS]
86 from [OTHER RESULTS]. It asks for Reflection, Research Plan and a Solution of five gene symbols
87 approved by the HUGO Gene Nomenclature Committee (HGNC), encouraging both observed hits
88 and exploration. Response formats and limits are in Table 1; up to three content attempts are allowed.
89 Qwen3 thinking is disabled. We do not fine-tune the acquisition LLMs on these screens. Parsing
90 normalizes symbols and rejects ineligible or repeated entries. If valid centres are missing, a fixed
91 hash ordering supplies replacements without using outcomes. We call this centre recovery. NN
92 expands centres by cosine similarity in normalized Achilles features, allocating 26/26/26/25/25 slots
93 and excluding previously tested or already selected genes. These centre and batch allocations specify
94 the tested procedure, not an established laboratory optimum.

95 A centre is an anchor, not an extra test. In the first planned Carnevale campaign using native JSON
96 output, 128 initial tests reveal six hits. Gemini returns ADORA2A, CERS4, POLR2F, PRKACA and
97 ZNF74; quotas 26/26/26/25/25 select 128 untested genes, excluding overlaps. Eight are released
98 hits, giving 14/256, not 14/261: centres do not add tests. An untested centre uses a quota slot if
99 selected. The source archive includes this exact prompt, response and batch replay; the example
100 illustrates accounting, not a typical effect. B160 uses quotas 7/7/6/6/6.

101 3.2 Matched interventions and rivals

102 The feedback interventions separate access to outcomes from correct assignment to genes. TRUE
103 presents correct gene–score–hit records. WITHIN permutes intact score/hit pairs among genes in
104 each acquired batch: outcome values stay the same, but their gene assignments change. NO re-
105 tains tested identities but withholds outcomes. Evaluation always uses the released hit labels. Only
106 acquired genes enter the feedback history, transformed according to TRUE, WITHIN or NO. Unac-
107 quired outcomes are used for evaluation after a batch has been selected; they are not available to its

controller or used to stop a campaign early. The future-label oracle in Section 4.5 is an explicitly separate, nondeployable analysis. The interventions measure behavioral use of supplied outcomes, not the faithfulness of the generated reasoning or its internal mechanism. Uniform-hit samples up to five distinct displayed hits; first-hit takes them in acquisition order. If fewer than five are available, both use the same outcome-blind centre completion as the LLM. Under NO, both select from tested genes without filtering on hit status. All share the executor. Random expansion retains centres that NN would include at the current state and fills remaining slots uniformly. The original prompts still describe NN: this executor cross compares complete policies under fixed instructions. Subsequent states may diverge.

Random acquisition samples untested genes uniformly. Coreset greedily chooses the gene whose nearest tested or newly selected neighbour is farthest away in cosine distance [Sener and Savarese, 2018]. Gaussian-process upper confidence bound (GP-UCB) and linear-UCB fit acquired outcomes. To choose successive genes within a batch before their results arrive, they provisionally substitute predicted outcomes [Srinivas et al., 2010, Li et al., 2010]. Among 20 settings on four separate random-gene development starts per screen, mean final hits select linear-UCB for Sanchez tau and GP-UCB otherwise (fixed tie breaking). CAR-T/ORCS1107 transfer the best cross-screen normalized-yield setting without retuning. Testing shares the LLM initial batch and features, unlike random-gene development; intervals condition on selection. This reference is not an optimized ceiling. Uniform-hit has no fitted parameters (Appendix I).

A *saved state* fixes the genes and outcomes already observed. Comparing actions there measures the next batch; completing their resulting histories measures final campaign yield. Multiple states from one start are related observations, not independent replications. External ORCS1107 is a primary-T-cell proliferation screen [Shifrut et al., 2018, Stark et al., 2006], with 17,883 eligible genes and 3,538 curated hits. ORCS1107 uses its curated HIT flag, matching released `Score.2 < 0.05` or `Score.4 < 0.05`, rather than the original panel’s top-mover lists. These source-specific criteria are not a common biological threshold.

3.3 Comparisons and uncertainty

Our comparisons ask three questions: does receiving outcomes help the LLM, does assigning them to the correct genes help, and does the LLM beat an informed numerical controller? For final hits Y , budget B and eligible released hits H , precision is Y/B and recall Y/H . Both rank policies by Y within a fixed readout/budget; pooling recall instead weights readouts by their different hit totals.

Let $Y_{p,a,e}$ denote final hits for controller p , feedback a and executor e . Here L/U are LLM/uniform-hit, $T/W/O$ are correct/permitted/withheld outcomes, and N/R are NN/random expansion. The three headline comparisons are

$$B_{\text{assign}} = \overline{Y_{L,T,N}} - \overline{Y_{L,W,N}}, \quad B_{\text{avail}} = \overline{Y_{L,T,N}} - \overline{Y_{L,O,N}}, \quad \Delta_{\text{NN}} = \overline{Y_{L,T,N}} - \overline{Y_{U,T,N}}.$$

Positive values favour the first policy in each subtraction. For one model–readout cell, we subtract paired yields at each of its 20 starts and average those differences; the bar then averages the 36 cell means equally. These paired campaign differences condition on the fixed panel, initialization and response format. Feedback changes later acquisitions and histories; its campaign effect does not isolate a decision at one fixed state.

The four prespecified contrasts are $\Delta_{\text{NN}}; \Gamma$, the LLM’s correct-minus-permuted gain minus uniform-hit’s corresponding gain; and Ω_L, Ω_U , each controller’s NN-minus-random expansion gain. Thus $\Omega_L - \Omega_U$ measures the change in the controller gap between executors (full equations: Appendix M). All contrasts match gene tests, not tokens, retries, API cost or wall-clock time.

We bootstrap paired starts within cells with 20,000 draws [Efron, 1979, Agarwal et al., 2021]. Bonferroni intervals are 98.75% for the four prespecified contrasts and 98.333% for the three retrospectively grouped headline contrasts. The latter reuse the same outcomes, not new confirmation. Follow-ups report separate families. Coverage is nominal under the bootstrap approximation (Appendix M).

These intervals quantify variation across computational starts on fixed screens, not variation across a population of biological sources [Dror et al., 2018, Demsar, 2006]. Correlated readouts remain grouped in source-weighting and source-omission checks. Zero-crossing intervals do not establish equivalence [Lakens, 2017].

4 Results

4.1 Feedback helps without establishing controller advantage

The LLM benefits from correctly assigned outcomes and from having outcomes available, yet uniform-hit recovers more hits. Table 2 jointly adjusts the three contrasts using the original saved outcomes (prespecified four-contrast family: Table 16).

Table 2: Three headline contrasts after 640 tests: 720 paired starts, 36 equally weighted model-readout cells, fixed biological sources. Retrospective Bonferroni 98.333% intervals provide nominal 95% family coverage.

Comparison under NN expansion	Mean hit difference	Adjusted interval
LLM: correct minus permuted outcomes	+6.88	[5.00, 8.73]
LLM: correct minus withheld outcomes	+18.18	[16.49, 19.84]
Correct outcomes: LLM minus uniform-hit	-9.98	[-11.40, -8.61]

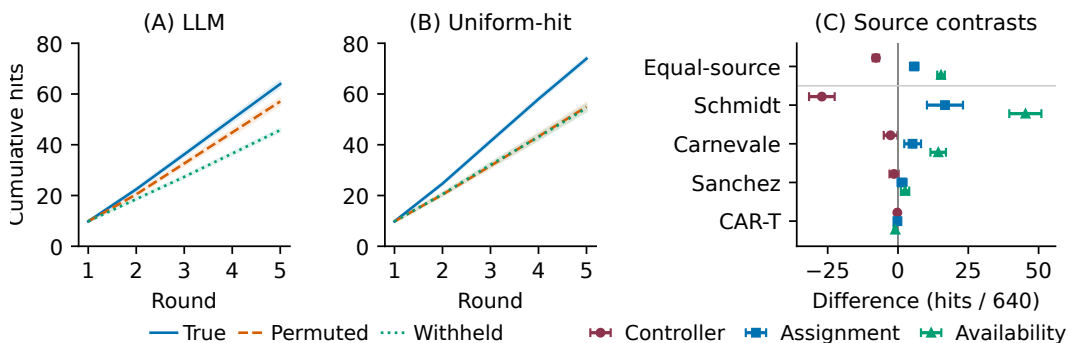


Figure 2: Feedback helps the LLM, but uniform-hit yields more hits. A–B: cumulative hits under NN expansion, equally weighted across 36 model-readout cells (720 starts); supporting 95% bands resample starts within cells. C: controller contrast is LLM minus uniform-hit. Assignment and availability subtract permuted and withheld yields, respectively, from the LLM’s correct-feedback yield. Intervals: 99.583% for twelve source contrasts, 98.333% for three equal-source contrasts.

Uniform-hit gains 18.94 hits from correct rather than permuted outcomes, compared with the LLM’s 6.88 (supporting 95% intervals [17.58,20.33] and [5.33,8.41]). The LLM-minus-first-hit difference is -1.28 [-2.64,0.06].

Absolute yields place these comparisons alongside random acquisition, Coreset and the development-selected classical reference (Figure 3). Appendix B.5 reports a separate classical-retuning sensitivity.

4.2 How representative is the pooled result?

A pooled loss can hide LLM-favouring conditions or be dominated by one source. All six model-level B640 controller contrasts are negative, ranging from -15.63 to -3.42, while 9 of 36 model-readout means favor the LLM. Omitting Schmidt reduces the pooled deficit from 9.98 to 1.43 hits. The original average weights readouts equally, so sources with more readouts contribute more. Equal weighting of the four sources gives -7.84 [-8.98,-6.71] (98.333%). The pooled magnitude therefore does not describe a typical independent source (Figure 4). Across starts the LLM wins 223, loses 460 and ties 37; the median paired difference is -3 hits. Scharenberg gives -2.94 [-3.33,-2.54] hits/160, with a post-hoc classical-reference advantage over uniform-hit.

The ORCS1107 external controller check also favors uniform-hit: -46.88 [-54.87,-38.90] hits/640. In the separate external feedback check, availability helps by 58.85 [32.20,82.60] hits/640, while the benefit of correct gene assignment remains unresolved at 10.98 [-15.72,35.15] (98.75%; Table 16).

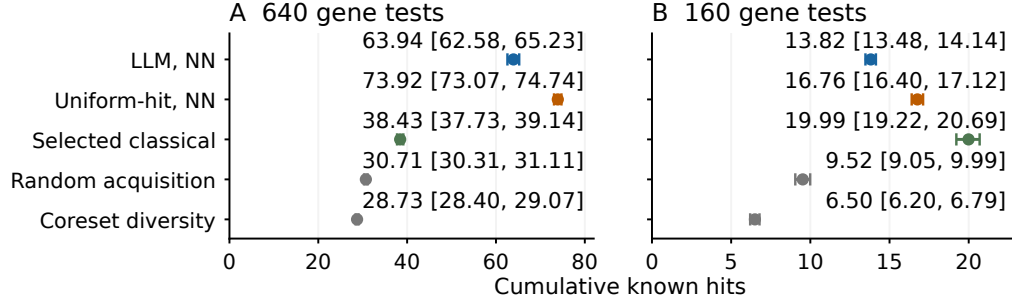


Figure 3: Absolute released-hit yield: means and supporting 95% intervals. A: six B640 readouts, 720 starts; B: Scharenberg B160, 120 starts. Screen and budget both differ. TRUE feedback where applicable. The classical reference is linear-UCB on Sanchez tau and GP-UCB on the other screens; each uses its own acquired outcomes. Interval overlap is not a paired significance test.

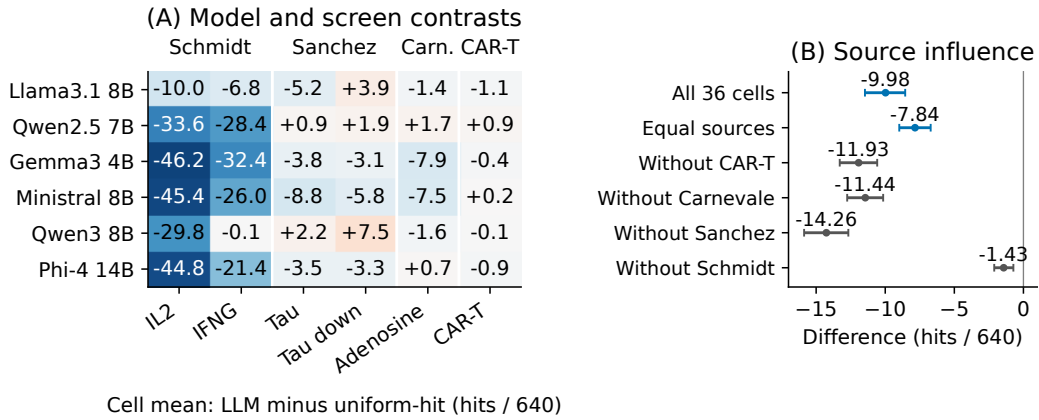


Figure 4: LLM-minus-uniform-hit differences by model and screen, and sensitivity to source weighting. A: all 36 TRUE/NN cell means (20 paired starts each); source boundaries group correlated readouts. Nine means favor the LLM and 27 uniform-hit; signs are descriptive. B: pooled 98.75%, equal-source 98.333% (as in Fig. 2C), and leave-source-out 95% intervals; these belong to different stated comparison families. Schmidt affects magnitude; all source omissions retain negative pooled intervals.

Original source-level joint checks support positive feedback benefits alongside a controller deficit on Schmidt and Carnevale; Sanchez is unresolved. CAR-T’s availability effect is negative, -0.96 $[-1.56, -0.36]$, while assignment and controller intervals cross zero.

4.3 Changing how centres become batches

The executor may help one controller more than the other, even when both use the same rule. NN expansion gains 32.27 hits for the LLM and 43.41 for uniform-hit. Switching to random filling reverses the pooled controller difference to $+1.16$ $[0.71, 1.63]$ hits/640; the controller-executor interaction is -11.14 $[-12.40, -9.94]$ (supporting 95%).

This is the change in the controller gap: $-9.98 - (+1.16) = -11.14$. NN improves both controllers, but improves uniform-hit more; the relative LLM win under random filling is not an absolute improvement in LLM yield. Within the random-expansion comparison, directly acquired centres contribute $+1.47$ hits and other genes -0.30 $[-0.74, 0.16]$. Uniform-hit normally proposes already tested hits, while the LLM can propose new genes. Together with the fixed NN-oriented instructions, this makes the reversal a complete-policy result rather than an isolated benefit of interpreting feedback. A separate fresh-start executor campaign also finds negative interactions for Qwen2.5 and Ministral (Appendix I).

The LLM’s next-batch deficit also remains when centres guide expansion but cannot be acquired. A separate exploratory test follows each controller’s batch with uniform-hit for the remaining rounds and also retains a deficit in final LLM-branch yield (Appendix A.2).

4.4 Initialization can reverse controller value

The preceding comparisons share an LLM-selected first batch. An informed rival may benefit from those starting genes even when it replaces all later LLM decisions. We therefore ask whether the controller comparison changes when the initial genes are selected differently. The initializer study (Table 1) uses Gemini3.8Flash and DeepSeekV4 Flash0731, with 40 fresh starts per source/model. Each of three starting rules is followed by either LLM or uniform-hit control; the initial 128 tests remain part of 640. Their JSON output formats differ from the original panel, so this comparison does not isolate model capacity.

The three initializers use random individual genes, random NN centres or LLM-selected NN centres. Comparing the last two holds neighbourhood expansion fixed. Comparing controllers after the same start measures continuation value; fixing uniform-hit continuation instead isolates starting-gene value. Their interaction asks whether an LLM-selected start changes the later LLM controller’s advantage:

$$I = \Delta_{\text{LLM-centre NN}} - \Delta_{\text{random-centre NN}}.$$

Each Δ is the controller gap after the named start. Negative I means LLM continuation fares relatively worse after LLM-selected centres. Both terms retain NN initialization; random individual genes are not this interaction’s reference. Figure 5 displays all three starting rules and their absolute outcomes.

After random-gene starts, DeepSeek beats uniform-hit and both deployments benefit from correct and available outcomes; Gemini’s controller direction is unresolved. Both pooled controller gaps are negative after LLM initialization. The interaction directly tests whether initialization changes the controller gap; it does not require both gaps to have individually resolved signs. Random-centre NN produces higher absolute means than LLM-centre NN for both continuations. Thus these results do not justify generally using the LLM to choose the first batch.

Table 3: Initializer–controller study: hits/640, paired 99.583% intervals (twelve tests; 80 starts/model, two sources). Controller: LLM-minus-uniform. Initializer value compares starting batches while keeping uniform-hit continuation fixed. Interaction compares the controller gaps after LLM-centre and random-centre NN initialization. Feedback contrasts: random-gene starts only.

Contrast	Gemini3.8Flash	DeepSeekV4
Controller after random genes	3.43 [-0.55,7.65]	5.11 [0.19,10.19]
Controller after LLM initialization	−22.86 [-33.90,-12.26]	−16.26 [-27.18,-6.37]
Outcome availability (random genes)	76.14 [70.36,80.98]	69.04 [63.33,74.41]
Correct assignment (random genes)	41.40 [30.43,51.92]	30.53 [19.96,41.85]
LLM minus random-centre initializer	−15.43 [-20.85,-9.84]	−10.86 [-15.62,-5.74]
Initializer × controller interaction	−24.54 [-36.00,-13.19]	−18.75 [-30.55,-7.86]

224

Gemini’s LLM-initialized contrast is source-dependent: −52.00 on IL2 and +6.28 on Carnevale (source intervals in Appendix D).

Gemini’s 1,680/1,680 valid decisions and DeepSeek’s 1,678/1,680 limit an incomplete-output explanation for this study, not for the original six-model panel. Feedback interventions were collected only after random-gene starts; their benefits cannot be combined with losses after LLM starts to claim both within one condition.

4.5 Selection gains do not establish an LLM-specific benefit

Can useful LLM batches be identified before testing, despite lower overall yield? The geometry selector compares NN-expanded LLM and uniform-hit batches by mean cosine distance from acquired non-centre genes to their selecting centres. It chooses the smaller mean; ties favor uniform-hit.

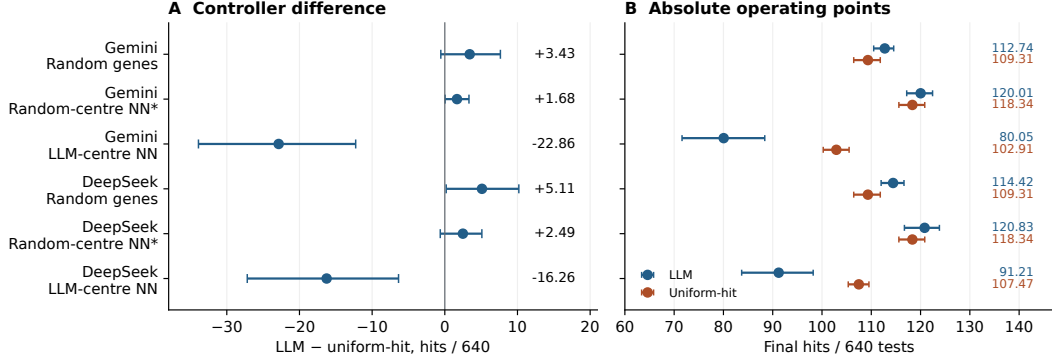


Figure 5: Initialization under correct feedback and NN continuation; 40 starts/source/model. Left: controller differences; 99.583% intervals for random-gene and LLM-centre starts, supporting 95% for random centres. Right: absolute yields with supporting 95% intervals. The interaction compares LLM-centre with random-centre NN, not random genes.

For the geometry selector, illustrative mean distances of 0.12 for the LLM proposal and 0.18 for uniform-hit select the LLM batch: its acquired neighbours lie closer to their anchors. This does not say they are hits; outcomes are still hidden.

We test the rule at the 720 original B640 first-adaptive states, then in full Llama3.1/Qwen2.5 campaigns: ten previously studied starts per model-readout cell give 120 B640 and 20 separate Scharenberg starts. The saved-state test fixes the observed history; full campaigns let histories diverge.

At the first adaptive decision, choosing the proposal with smaller centre-to-neighbour distance gains 0.46 [0.26,0.67] hits/128 over always choosing uniform-hit. In full Llama/Qwen2.5 campaigns the geometry selector gains 2.78 [1.78,3.83] hits/640 over uniform-hit. This comparison gives the selector two proposals but uniform-hit only one. We therefore also give the same selector two independent uniform-hit proposals, to isolate the value of including an LLM proposal. The mixed-minus-two-uniform difference is -0.11 $[-1.32,1.08]$ (primary 97.5%); this comparison does not establish an LLM-specific benefit.

Candidate supply versus selection. We also separate useful proposals from the selector’s ability to choose them. At all four decisions of the 120 mixed-selector histories (480 states), we compare the LLM-plus-uniform pool with two uniform proposals at the same observed state. A hypothetical oracle chooses using future hit labels, which a real selector cannot see. The mixed pool offers it $+0.43$ $[0.19,0.68]$ hits/128 over two uniform proposals; geometry realizes an unresolved $+0.15$ $[-0.14,0.43]$. These exploratory 95% intervals resample starts with their four decisions together. Local opportunities cannot be summed into campaign gains because changing a batch also changes subsequent histories. The LLM sometimes supplies a better batch, but the tested rule has not established a gain over selecting between two numerical proposals (Appendix A.2).

4.6 Response interfaces and prior exposure

We distinguish poor gene choices from failure to return all five valid centres. The original TRUE/NN population includes 162/720 campaigns with at least one incomplete adaptive proposal. The stricter sensitivity applies successive requirements: 717 starts have complete initialization, 557 also have complete TRUE/NN proposals, 502 also have complete WITHIN/NN proposals, and 476 also have complete TRUE/random proposals. This is why the final subset is not simply $720 - 162$. It retains a controller deficit, -9.23 $[-10.79,-7.63]$ hits/640 on 476 starts. This selected subset cannot show how omitted campaigns would perform without recovery; finite-pool bounds allow either ranking (Appendix B). Separate one-step tests improve completion at 1,536 tokens yet retain a -1.63 $[-2.24,-1.05]$ LLM-minus-uniform difference per 128 tests. Gene-name masking, Phi-4 exclusion and full recovery accounting remain in Appendices H and B; none substitutes for a complete campaign.

Complete campaigns with expanded response formats. The expanded-response and native-response studies in Table 1 repeat all three contrasts. Qwen2.5/Minstral use solution-only lists

with strict normalization; Gemini3.8Flash uses native JSON and low reasoning. Each has twenty starts per model–source cell on Carnevale/IL2 and previously saved LLM initial batches. All four adaptive rounds and uniform-hit controls are generated afresh, with each policy receiving the outcomes of its own earlier choices. These are new decisions on previously studied screens and starts, not new biological validation or an isolated token-budget intervention (Appendix A.3).

Within each panel, feedback and controller comparisons share interface and initialization. Both retain positive feedback benefits but negative pooled controller contrasts (Figure 6). LLM-minus-uniform-hit is -8.35 $[-14.21, -3.25]$ hits/640 in the small-model panel and -11.45 $[-24.94, -0.33]$ for Gemini (separate three-test 98.333% families). Gemini supplies five valid centres in all 480 adaptive decisions: 40 starts $\times$ three feedback arms $\times$ four rounds, without recovery. The independent start count remains 40. Its controller contrast is nevertheless positive on Carnevale, $+6.85$ $[4.05, 9.65]$, and negative on IL2, -29.75 $[-51.25, -11.10]$ (supporting 95%). Thus output truncation or recovery is not necessary for the observed distinction, while source conditions still change which controller wins.

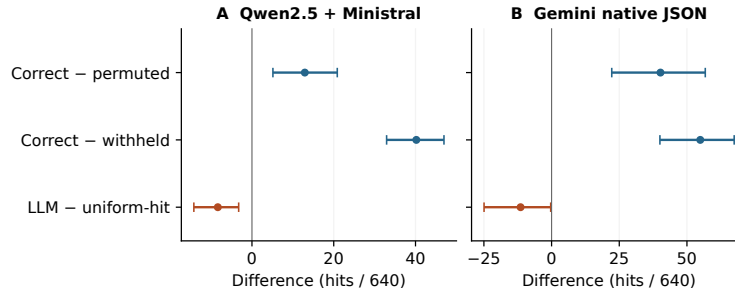


Figure 6: Feedback benefits and controller differences in complete campaigns. Each panel compares correct with permuted outcomes, correct with withheld outcomes, and LLM with uniform-hit, in the same row order. Separate three-test 98.333% intervals; sample sizes in Table 1. Gemini’s pooled loss includes a Carnevale win and IL2 loss.

283

Published screens and released CAR-T labels may have entered training data; checkpoint-specific exposure is unknown. Publication status, biology and feature geometry are confounded, and both policies inherit the LLM-selected initial batch. Neither the CAR-T comparison nor name masking certifies contamination-free evaluation (Appendix B).

288 5 Discussion

The numerical rival retains outcomes and batch expansion; the two-proposal control retains the choice between alternatives. These controls separate feedback use from added LLM yield. The source and initializer reversals remain unexplained. We asked whether the observed history predicts the LLM’s next-batch advantage. On 720 first adaptive decisions, we added three features to hit count: how much of the untested pool neighbours known hits, how often those neighbourhoods overlap, and how much of their original neighbourhoods is already tested. Prediction error fell on only one of four sources held out from fitting (Appendix E.2).

The test separating executor instructions from permission to propose new centres remains incomplete (18/20 starts). Accounting for unfinished campaigns and variation across starts leaves its contrasts unresolved (Appendix H.5). Only Achilles features were tested; released labels, unknown pretraining exposure and shared biology limit transport. No human comparison or prospective efficacy was measured. Failed horizon predictions remain in Appendix J.

301 6 Conclusion

Feedback benefits can coexist with lower yield than informed numerical control. Source and initializer reversals prevent a universal ranking: test continued-control value on the intended screen. Retrospective hits establish neither laboratory benefit nor end-to-end savings.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, et al. Phi-4 Technical Report. *arXiv preprint arXiv:2412.08905*, 2024. URL <https://arxiv.org/abs/2412.08905>.
- Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In *Advances in Neural Information Processing Systems*, 34, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html>.
- Adam Bayley, Xiaodan Zhu, Raquel Aoki, Yanshuai Cao, and Kevin H. Wilson. Jump start or false start? a theoretical and empirical evaluation of llm-initialized bandits. *Transactions on Machine Learning Research*, 2026. URL <https://openreview.net/forum?id=tojKjqIOBd>.
- BioDiscoveryAgent contributors. Achilles feature data distributed through the biodiscoveryagent downloader, n.d. URL <https://figshare.com/ndownloader/files/49843176>.
- Broad Institute. Depmap portal: Data resources, n.d. URL https://depmap.org/portal/data_page/.
- Julia Carnevale, Eric Shifrut, Nupura Kale, William A. Nyberg, Franziska Blaeschke, Yan Yi Chen, Zhongmei Li, Sagar P. Bapat, Morgan E. Diolaiti, Patrick O’Leary, et al. RASA2 ablation in T cells boosts antigen sensitivity and long-term function. *Nature*, 609(7925):174–182, 2022.
- Hyungjoo Chae, Namyoung Kim, Kai Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 63707–63738, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a00548031e4647b13042c97c922fadb1-Paper-Conference.pdf.
- Janez Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006. URL <https://jmlr.org/papers/v7/demsar06a.html>.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, 2018. doi: 10.18653/v1/P18-1128. URL <https://doi.org/10.18653/v1/P18-1128>.
- Carl Edwards, Edward De Brouwer, Xiner Li, Namkyeong Lee, Ehsan Hajiramezanali, Anne Biton, Sara Mostafavi, and Gabriele Scalia. Biology-in-the-loop: Amortized adaptive hit discovery in crispr screens. *arXiv preprint arXiv:2609.11877*, 2026. URL <https://arxiv.org/abs/2609.11877>.
- B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1979. doi: 10.1214/aos/1176344552. URL <https://doi.org/10.1214/aos/1176344552>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Rushil Gupta, Jason Hartford, and Bang Liu. LLMs for Bayesian optimization in scientific domains: Are we there yet? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15482–15510. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.838. URL <https://aclanthology.org/2025.findings-emnlp.838/>.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html.

352 Agustinus Kristiadi, Felix Strieth-Kalthoff, Marta Skreta, Pascal Poupart, Alan Aspuru-Guzik, and
353 Geoff Pleiss. A Sober Look at LLMs for Material Discovery: Are They Actually Good for
354 Bayesian Optimization Over Molecules? In *Proceedings of the 41st International Conference on*
355 *Machine Learning*, 2024. URL [https://proceedings.mlr.press/v235/kristiadi24a.](https://proceedings.mlr.press/v235/kristiadi24a.html)
356 [html](https://proceedings.mlr.press/v235/kristiadi24a.html).

357 Daniel Lakens. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-
358 Analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/
359 1948550617697177. URL <https://doi.org/10.1177/1948550617697177>.

360 Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to
361 personalized news article recommendation. In *Proceedings of the 19th International Conference*
362 *on World Wide Web*, pages 661–670, 2010. doi: 10.1145/1772690.1772758.

363 Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models
364 to enhance bayesian optimization. In *The Twelfth International Conference on Learning Repre-*
365 *sentations*, 2024. URL <https://openreview.net/forum?id=00xotBmGol>.

366 Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and
367 Patrick Schwab. Genedisco: A benchmark for experimental design in drug discovery. In
368 *International Conference on Learning Representations*, 2022. URL [https://github.com/](https://github.com/genedisco/genedisco)
369 [genedisco/genedisco](https://github.com/genedisco/genedisco).

370 Robin M Meyers, Jordan G Bryan, James M McFarland, Barbara A Weir, Ann E Sizemore, Han
371 Xu, et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9
372 essentiality screens in cancer cells. *Nature Genetics*, 49(12):1779–1784, 2017. doi: 10.1038/ng.
373 3984. URL <https://doi.org/10.1038/ng.3984>.

374 Mistral AI. Mistral 3 8B Instruct 2512: Model Card. *Model documentation*, 2025. URL [https:](https://huggingface.co/mistralai/Mistral-3-8B-Instruct-2512)
375 [//huggingface.co/mistralai/Mistral-3-8B-Instruct-2512](https://huggingface.co/mistralai/Mistral-3-8B-Instruct-2512).

376 Giang Ngo, Dat Phan Trong, Dang Nguyen, Sunil Gupta, and Svetha Venkatesh. Adaptive Ac-
377 quisition Selection for Bayesian Optimization with Large Language Models. In *International*
378 *Conference on Learning Representations*, 2026. URL [https://openreview.net/forum?id=](https://openreview.net/forum?id=EPKmSgXvRe)
379 [EPKmSgXvRe](https://openreview.net/forum?id=EPKmSgXvRe).

380 Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 Technical Report.
381 *arXiv preprint arXiv:2412.15115*, 2024. URL <https://arxiv.org/abs/2412.15115>.

382 Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexan-
383 der Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI agent for de-
384 signing genetic perturbation experiments. In *International Conference on Learning Repre-*
385 *sentations*, 2025. URL [https://proceedings.iclr.cc/paper_files/paper/2025/hash/](https://proceedings.iclr.cc/paper_files/paper/2025/hash/4252dc94531833029000f85dc5fac792-Abstract-Conference.html)
386 [4252dc94531833029000f85dc5fac792-Abstract-Conference.html](https://proceedings.iclr.cc/paper_files/paper/2025/hash/4252dc94531833029000f85dc5fac792-Abstract-Conference.html).

387 Carlos G. Sanchez, Christopher M. Acker, Audrey Gray, Malini Varadarajan, Cheng Song, Nadire R.
388 Cochran, Steven Paula, Alicia Lindeman, Shaojian An, Gregory McAllister, et al. Genome-wide
389 CRISPR screen identifies protein pathways modulating tau protein levels in neurons. *Communi-*
390 *cations Biology*, 4(1):736, 2021.

391 Samantha G. Scharenberg, Wentao Dong, Ali Ghoochani, Kwamina Nyame, Roni Levin-
392 Konigsberg, Aswini R. Krishnan, et al. An SPNS1-dependent lysosomal lipid transport path-
393 way that enables cell survival under choline limitation. *Science Advances*, 9(16), 2023. doi:
394 10.1126/sciadv.adf8966. URL <https://doi.org/10.1126/sciadv.adf8966>.

395 Ralf Schmidt, Zachary Steinhart, Madeline Layeghi, Jacob W. Freimer, Raymund Bueno, Vinh Q.
396 Nguyen, Franziska Blaesche, Chun Jimmie Ye, and Alexander Marson. CRISPR activation and
397 interference screens decode stimulation responses in primary human T cells. *Science*, 375(6580):
398 eabj4008, 2022. doi: 10.1126/science.abj4008.

399 Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set
400 approach. In *International Conference on Learning Representations*, 2018. URL [https://](https://openreview.net/forum?id=H1aIuk-RW)
401 openreview.net/forum?id=H1aIuk-RW.

402 Eric Shifrut, Julia Carnevale, Victoria Tobin, et al. Genome-wide crispr screens in primary human t
403 cells reveal key regulators of immune function. *Cell*, 175(7):1958–1971.e15, 2018. doi: 10.1016/
404 j.cell.2018.10.024.

405 Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian optimization of machine
406 learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

407 Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian Process Op-
408 timization in the Bandit Setting: No Regret and Experimental Design. In *Proceedings of the 27th*
409 *International Conference on Machine Learning*, 2010. URL [https://icml.cc/Conferences/](https://icml.cc/Conferences/2010/papers/422.pdf)
410 [2010/papers/422.pdf](https://icml.cc/Conferences/2010/papers/422.pdf).

411 Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike
412 Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34
413 (suppl_1):D535–D539, 2006. doi: 10.1093/nar/gkj109.

414 Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limita-
415 tions of large language models on reasoning and planning tasks. In *International Conference*
416 *on Learning Representations*, 2025. URL [https://proceedings.iclr.cc/paper_files/](https://proceedings.iclr.cc/paper_files/paper/2025/hash/f3c5e56274140e0420baa3916c529210-Abstract-Conference.html)
417 [paper/2025/hash/f3c5e56274140e0420baa3916c529210-Abstract-Conference.html](https://proceedings.iclr.cc/paper_files/paper/2025/hash/f3c5e56274140e0420baa3916c529210-Abstract-Conference.html).

418 Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej,
419 et al. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025. URL [https://](https://arxiv.org/abs/2503.19786)
420 arxiv.org/abs/2503.19786.

421 Aviad Tsherniak, Francisca Vazquez, Phil G. Montgomery, Barbara A. Weir, Gregory Kryukov,
422 Glenn S. Cowley, et al. Defining a Cancer Dependency Map. *Cell*, 170(3):564–576.e16, 2017.
423 doi: 10.1016/j.cell.2017.06.010. URL <https://doi.org/10.1016/j.cell.2017.06.010>.

424 Gilles Wainrib, Barbara Bodinier, Haitem Dakhli, Josep Monserrat, Almudena Espin Perez, Sab-
425 rina Carpentier, Roberta Codato, and John Klein. Can AI scientist agents learn from lab-in-the-
426 loop feedback? evidence from iterative perturbation discovery. *arXiv preprint arXiv:2603.26177*,
427 2026. URL <https://arxiv.org/abs/2603.26177>.

428 Haorui Wang, Marta Skreta, Cher Tian Ser, Wenhao Gao, Lingkai Kong, Felix Strieth-Kalthoff,
429 Chenru Duan, Yuchen Zhuang, Yue Yu, Yanqiao Zhu, Yuanqi Du, Alan Aspuru-Guzik, Kirill
430 Neklyudov, and Chao Zhang. Efficient evolutionary search over chemical space with large lan-
431 guage models. In *The Thirteenth International Conference on Learning Representations*, 2025.
432 URL <https://openreview.net/forum?id=awWiNvQwf3>.

433 An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report. *arXiv preprint*
434 *arXiv:2505.09388*, 2025a. URL <https://arxiv.org/abs/2505.09388>.

435 Ke Yang, Yao Liu, Sapana Chaudhary, Rasool Fakoor, Pratik Chaudhari, George
436 Karypis, and Huzefa Rangwala. AgentOccam: A simple yet strong baseline for
437 LLM-based web agents. In *International Conference on Learning Representations*,
438 2025b. URL [https://proceedings.iclr.cc/paper_files/paper/2025/hash/](https://proceedings.iclr.cc/paper_files/paper/2025/hash/f2c6e459b95694a24ac69c469a4ee746-Abstract-Conference.html)
439 [f2c6e459b95694a24ac69c469a4ee746-Abstract-Conference.html](https://proceedings.iclr.cc/paper_files/paper/2025/hash/f2c6e459b95694a24ac69c469a4ee746-Abstract-Conference.html).

440 Xinzhe Yuan, Zhuo Chen, Jianshu Zhang, Huan Xiong, Nanyang Ye, Yuqiang Li, and Qinying
441 Gu. Unleashing LLMs in bayesian optimization: Preference-guided framework for scientific
442 discovery. In *The Fourteenth International Conference on Learning Representations*, 2026. URL
443 <https://openreview.net/forum?id=LktU0ZayG9>.

444 **Supplementary disclosures**

445 **AI use**

446 Generative AI tools, including Codex, assisted with literature search and comparison, refinement of
447 hypotheses and experimental protocols, implementation and debugging, data preparation and analy-
448 sis, interpretation, figure generation, and manuscript drafting and revision. Synthetic feedback was
449 used only for technical interface pilots and is separated from biological evaluation. Study outcomes
450 come from released screen records; language-model outputs are the evaluated acquisition decisions.
451 Saved raw responses, deterministic replay and numerical consistency checks support verification of
452 the AI-assisted work. The authors are responsible for the final claims and artifacts.

453 **Reproducibility**

454 The main methods specify the acquisition contract, outcome definitions, matched controls and sta-
455 tistical units. The technical appendix records model interfaces, experiments and preserved failures.
456 The appendix specifies source contracts and the distinctions among saved-evidence analysis, deter-
457 ministic replay and paid stochastic regeneration. Upstream inputs with unresolved redistribution
458 rights are identified by source and checksum rather than redistributed. Hosted model versions may
459 change, so exact saved-output reproduction is distinguished from fresh sampling. The accompa-
460 nying source archive replays saved outcomes and audits confirmation histories and responses. The
461 worked example reconstructs a batch from frozen neighbour ranks; these checks do not regenerate
462 all-pool neighbours or model calls. Public hosting and feature access remain unresolved.

463 **Guide to the supplementary material**

464 The appendix supports the main comparisons and their alternatives: task and candidate-selection
465 checks (Appendix A); recovery and reference policies (Appendix B); complete acquisition contracts
466 (Appendix C); initialization (Appendix D); original and external panels (Appendices F, G); inter-
467 face and executor controls (Appendices H, I); and inference definitions (Appendix M). Source and
468 reproduction boundaries are in Appendices K.1 and K.3.

469 Earlier unsuccessful predictions are summarized in Appendix J. The preserved full version contains
470 their complete tables and additional annotation-diversity and feature-clustering explorations. Those
471 explorations found objective-dependent rankings: equal-test annotation coverage favored uniform-
472 hit, whereas equal-hit composition and some GO follow-up objectives favored LLM hit sets. Feature
473 clustering did not establish a causal explanation or a forward controller-selection rule. These sup-
474plementary explorations are not needed to establish the main paper's hit-yield claims and do not
475 provide additional biological replications. Their full positive and negative results remain preserved
476 together.

A Evidence map and additional analyses

The main comparisons concern different decision problems. Table 4 separates their populations. Original, retrospective and follow-up uncertainty families remain distinct; none represents biological-source sampling uncertainty.

Table 4: Questions, experimental units and scope. Starts pair campaigns through a shared initial batch; states fix one observed history. Repeated decisions do not add sources. Follow-ups largely reuse exposed conditions.

Question / role	Population and unit	Compared outcomes
Original feedback/control	Six models; six readouts/four sources; 720 B640 starts	Final hits: feedback and informed controller contrasts
Source and budget checks	Scharenberg: 120 B160 starts. ORCS1107: 120 controller starts; separate 40-start feedback study	Final hits within each study; B160 and B640 kept separate
Downstream value (exploratory)	80 states/40 starts, two models/two sources; 1,280 paired-seed branches	Batch and final hits under uniform-hit continuation
Initialization dependence	Gemini/DeepSeek, Carnevale/IL2; 80 starts/model; three initializers	Final hits: controller, initializer and their interaction
Campaign selection	Llama/Qwen2.5; 120 B640 and separate 20 B160 reused starts	Final hits: mixed proposals versus one or two uniform proposals
Candidate supply: exploratory	480 decisions/120 mixed-selector B640 starts	Batch hits: future-label oracle versus geometry rule
Batch preview: exploratory	40 fixed states/20 prior starts; two selectors, three information views	Batch hits: three information views
Policy prediction (exploratory)	720 original B640 starts; source and model-family holdout	Prediction error and final policy-selection yield
Response interfaces	Qwen2.5/Minstral: 80 starts; Gemini: 40; Carnevale/IL2, saved initial batches and fresh adaptive histories	Final hits: feedback and controller contrasts per interface

A.1 Task definitions and evaluated outcomes

The task description supplied to the controller and the labels used to score its acquired genes define the evaluation. All 25,339 original-panel requests were checked against source task and measurement strings, including prompt hashes. All 4,200 TRUE/NN LLM batches retain eligible, unique acquired genes, correct accumulated histories and released-label/score lookup. The literal Sanchez tau task is bidirectional, whereas tau-down is decrease-only. Both processed score files contain negative values; their sign alone does not identify the biological direction of a hit. Exact Schmidt assay-variant mapping remains unresolved. These checks establish the experiment’s mapping to its released benchmark, not independent biological reannotation.

Table 5: Task-to-outcome mapping. Eligible populations intersect released scores with Achilles features. Hits are inherited array memberships; different readouts from one source are not independent sources.

Readout	Genes	Hits	Literal task target
Carnevale adeno-sine	17,136	858	upon being knocked out, would boost the efficacy of engineered T cells in the presence of an adenosine agonist that creates an immunosuppressive condition
Sanchez tau	17,204	859	when knocked out, either increase or decrease expression of endogenous tau protein levels in neurons
Schmidt IL2	17,758	634	regulate the production of Interleukin-2 (IL-2)
Schmidt IFNG	17,009	861	regulate the production of Interferon-gamma (IFNG)

Readout	Genes	Hits	Literal task target
Sanchez tau-down	17,204	883	when knocked out, decrease expression of endogenous tau protein levels in neurons
Scharenberg	1,013	48	identify genes mediating lysosomal choline recycling using an endolysosome-focused CRISPR-Cas9 screen
CAR-T	17,608	139	upon activation allow cells to resist T-cell exhaustion, under the HA GD2 CAR (chimeric-antigenic receptor) condition

The source package’s `synthesis_reproduction/TASK_AUDIT.json` binds every task, score file and hit array by SHA256. `CURRENT_PROMPT_EXAMPLES.json` retains actual initialization and first-adaptive request payloads for each readout, with their original wording, prompt hashes, model/provider settings and token limits. These are current-panel interfaces; Appendix K.2 displays an actual initialization and adaptive example. No corrected wording is retroactively substituted into recorded requests.

The primary campaign recovery census is 162/720 starts with an incomplete adaptive TRUE/NN proposal. The 476-start complete mask additionally requires a complete shared initialization and every adaptive proposal in WITHIN/NN and TRUE/random; it does not require NO. The source package provides this exact mask, separate from the decision-level census.

A.2 Saved-candidate quality and common continuation

These are retrospective analyses of exposed sources and saved model decisions, frozen before their new calculations. They do not add independent biological validation or test an unobserved LLM response. All intervals below are exploratory 95%, conditional on the specified fixed sources, states and numerical seeds; 20,000 stratified paired bootstrap draws resample starts within model-readout cells and retain repeated decisions together.

Common continuation. For Qwen2.5 and Ministral on Carnevale and IL2, we use original starts 5001–5010 at adaptive rounds 1 and 3. Each branch preserves the complete acquired prefix. The saved LLM proposal is compared with a uniform-hit proposal at that identical state; each branch then runs uniform-hit on its own new history to the original 640-test budget. Eight paired numerical seeds per state yield 1,280 suffixes. Seed means are formed before start resampling; intervals do not include additional uncertainty from sampling another set of continuation seeds. A new history never receives an old LLM suffix. Positive subsequent compensation need not erase an immediate loss.

Table 6: LLM-batch minus uniform-batch value on identical saved states. Immediate yield counts the next 128 tests; continuation value also counts the remaining uniform-hit suffix. Compensation is their difference. 40 starts per round pooled, 20 per source; eight numerical seeds each.

Round	Source	Immediate	With continuation	Compensation
1	Pooled	−7.50 [−10.38,−4.45]	−6.27 [−9.42,−3.11]	+1.23 [−0.60,2.95]
1	IL2	−16.21 [−21.79,−10.30]	−13.51 [−19.52,−7.58]	+2.70 [−0.48,5.56]
1	Carnevale	+1.21 [−0.34,2.81]	+0.97 [−0.91,2.88]	−0.24 [−2.05,1.59]
3	Pooled	−6.42 [−9.41,−3.56]	−5.71 [−8.10,−3.43]	+0.71 [−0.25,1.64]
3	IL2	−11.18 [−16.96,−5.68]	−10.04 [−14.64,−5.74]	+1.14 [−0.66,2.88]
3	Carnevale	−1.67 [−2.88,−0.17]	−1.38 [−2.65,−0.01]	+0.29 [−0.34,0.94]

The pooled deficit persists after this common continuation; the first-round Carnevale contrast remains unresolved, whereas IL2 carries the large loss. The compensation intervals cross zero. Neither result establishes a universal absence of exploratory value or a causal mechanism for full LLM campaigns. Cross-history LLM responses would require new model calls.

Candidate supply. We inspect both candidates at every logged mixed-selector state: Llama3.1/Qwen2.5, ten starts per readout, four adaptive rounds, seven readouts. There are 480

B640 and 80 separate B160 states. At each, generate another uniform-hit candidate using the original SECOND seed rule. The mixed pool is the saved LLM and first uniform candidate; its equal-count rival is the first and second uniform candidates at that same state. Future-label oracle yield is the maximum candidate hit count, and regret is that maximum minus geometry-selected yield. The selection rule sees distances only. Budgets, states and candidate count are matched; candidate outcomes are analyst-visible diagnostics, not feedback supplied to a deployed policy.

Table 7: Mean differences and regret per decision, with starts as the resampling unit. Oracle increment is $\max(\text{LLM}, \text{U1}) - \max(\text{U1}, \text{U2})$; selected increment uses geometry instead of future labels. B160 changes the source and budget together.

Quantity	B640: 120 starts	B160: 20 starts
Mixed-minus-two-uniform oracle	+0.43 [0.19,0.68]	+0.11 [-0.11,0.34]
Mixed-minus-two-uniform selected	+0.15 [-0.14,0.43]	+0.05 [-0.15,0.24]
Mixed-pool regret	+1.17 [0.97,1.39]	+0.38 [0.20,0.56]
Two-uniform regret	+0.89 [0.71,1.09]	+0.31 [0.14,0.56]
Mixed oracle minus one uniform	+1.83 [1.60,2.08]	+0.39 [0.20,0.60]
Two-uniform oracle minus one	+1.40 [1.18,1.64]	+0.28 [0.15,0.44]

On B640, mean centre-set Jaccard overlap is 0.027 for LLM versus uniform-hit; acquired-batch overlap is 0.268, versus 0.558 between two uniform batches. Execution changes the alternatives, but does not erase all useful LLM candidate opportunity. Oracle headroom exceeds selected gain; this motivates testing selection only under held-out development and equal candidate counts, not assuming that a useful selector already exists. A separate complete-policy oracle on the original 720 starts gains +2.04 [1.75,2.36] hits/640 over uniform-hit. It chooses using final outcomes and is neither a pre-call gate nor an online switching policy.

Which available candidate does geometry choose? A further retrospective audit reconstructs the recorded geometry decisions from their centre-to-neighbour distances and binds the selected batches to their saved hit counts. Table 8 separates beneficial choices, harmful choices and hit ties. These are states on the previously collected mixed-selector trajectories, not new starts or independent biological replicates. The audit changes neither the candidates nor the selector and introduces no new model calls.

Table 8: Saved mixed-pool choices, by the true ordering of the two offered batches. Each row counts states, not independent starts. B640 has 480 states from 120 starts; B160 has 80 states from 20 starts and is kept separate. Columns identify the candidate selected by geometry. Hit ties are retained and are not counted as correct or incorrect strict choices.

Candidate ordering	B640		B160	
	LLM	Uniform	LLM	Uniform
LLM has more hits	68	149	5	15
Uniform has more hits	22	178	3	42
Same number of hits	16	47	2	13

On B640, geometry chooses a higher-yield candidate in 246/417 non-tied mixed comparisons, 59.0% [54.4,63.5]. For the two-uniform pool it chooses a higher-yield candidate in 224/367 non-tied comparisons, 61.0% [56.3,65.8]. These conditional rates have different eligible sets and are not a paired comparison of selection accuracy. The hit cost of a mistake also varies across states. We therefore compare lost hits on the same states: regret is the best available candidate's yield minus the selected yield, and the paired mixed-minus-two-uniform regret is +0.279 [0.006,0.556] hits per 128-test decision. It accounts for the difference between the oracle and selected increments because, state by state, selected increment equals oracle increment minus excess regret.

The new intervals are exploratory 95% percentile intervals from 20,000 bootstrap draws, seed 20260924, resampling whole starts within model-readout cells and retaining all four rounds and both candidate pools together. Conditional rates are recomputed within each draw. This new seed is used for this extension; the preceding table retains its original estimates and intervals. B160 excess

regret is +0.063 [-0.250,0.338] hits per 32-test decision and remains unresolved. The near-zero B640 lower endpoint is not a confirmatory mechanism test. Neither these local misses nor the oracle establish that another implementable selector improves complete-campaign yield. Portable records and calculations are supplied in `selection_reproduction/`.

Numerical replay boundary. Float32 cosine near-ties can exchange neighbours or their acquisition order. We retain recorded proposals and outcomes for the primary audit, validate released labels exactly, and verify each discrepant saved greedy step against its own available pool within 5×10^{-7} cosine similarity. Recomputing all three candidate batches on the current backend leaves the B640 oracle and selected increments unchanged. Full discrepancy records, including altered gene sets, hit counts and technical-stop amendments, accompany the data; this is not a claim of bit-identical full-pipeline regeneration.

Existing endpoint and distribution checks. The original controller difference under the stored alternative $|z| \geq 2$ endpoint is -9.98 [-11.12,-8.89]hits/640 (supporting 95% replay), rather than a new biological validation. Median released-hit difference is -3, with quartiles -15 and +2; the original win/loss/tie census remains 223/460/37. Source-specific absolute, initial and adaptive means and the complete distribution are retained in the source package.

A.3 Full-campaign response-contract checks

The two panels use different previously saved LLM initializations on the same exposed sources, Carnevale and Schmidt IL2. They are analyzed separately. Every continuation acquires four new 128-gene batches after the shared initial 128, for 640 tests. All adaptive histories are newly generated, including the uniform-hit reference. No old LLM suffix is inserted into a changed history. The small-model panel has 20 starts per model-source cell (80 total); the Gemini panel has 20 per source (40 total). Each panel compares LLM correct, tuple-permuted and withheld outcomes, and uniform-hit under correct outcomes. Evaluation always uses released hit labels.

Qwen2.5/Minstral use temperature 0.2 and a 1,536-token, solution-only contract. Strict bare symbol lists are normalized when the Solution heading is absent; prose is not mined for genes. The original eligible-pool and duplicate checks, three content attempts, deterministic recovery and NN executor remain. Gemini3.8Flash uses Google AI Studio, temperature 1, low reasoning, a 4,096-token allowance and native JSON constrained to a five-string centres array. The prompt explicitly permits tested centres, excludes repeated acquisitions and asks for cumulative hit yield. The parser, output budget, wording and serving epoch differ across studies: their historical comparison is not an isolated token-budget or model-capacity effect.

The four-block small-model pilot was excluded before aggregate biological outcomes were examined: its parser rejected bare symbol lists, and one route encountered rate limits. The reported panel uses 80 newly started blocks with identifiers fixed before collection. Numerical verification uses a 10^{-12} tolerance for derived absolute z-scores (observed discrepancies approximately 10^{-16}); raw scores, hit labels and gene identities match exactly. Failed calls are retained and conservatively charged. Transport failures are not counted as completed trajectories.

Table 9: New complete-campaign comparisons, hits per 640 tests. Each panel has its own three-test Bonferroni 98.333% family. Means equally weight model-source cells (small models) or sources (Gemini). Intervals resample starts within cells, keeping all arms paired, with 20,000 draws. They condition on the two fixed sources and saved initial batches.

Comparison	Qwen2.5/Minstral	Gemini3.8Flash
Correct minus permuted	+12.93 [+5.13, +20.87]	+40.23 [+22.23, +56.78]
Correct minus withheld	+40.20 [+32.93, +46.96]	+54.93 [+40.02, +67.45]
LLM minus uniform-hit	-8.35 [-14.21, -3.25]	-11.45 [-24.94, -0.33]

Five legal controller-supplied centres are obtained after allowed content attempts in 302/320, 315/320 and 317/320 small-model decisions under correct, permuted and withheld outcomes, respectively. Recovery supplies 21, five and three centres. Gemini completes 160/160 decisions in each arm, without recovery or content retries. These are adaptive-decision counts, not counts of

Table 10: Supporting source-specific contrasts, hits/640 with 95% paired intervals. The small-model cells pool two model families; Gemini is a separate deployment/population.

Panel	Source	Contrast	Estimate [95% interval]
Small models	Carnevale	Correct-permuted	+6.33 [+3.08, +9.88]
Small models	Carnevale	Correct-withheld	+10.10 [+8.22, +12.18]
Small models	Carnevale	LLM-uniform	+1.40 [-0.78, +3.60]
Small models	IL2	Correct-permuted	+19.53 [+7.13, +32.10]
Small models	IL2	Correct-withheld	+70.30 [+59.00, +81.05]
Small models	IL2	LLM-uniform	-18.10 [-27.15, -9.88]
Gemini	Carnevale	Correct-permuted	+16.60 [+8.95, +24.55]
Gemini	Carnevale	Correct-withheld	+26.60 [+24.70, +28.50]
Gemini	Carnevale	LLM-uniform	+6.85 [+4.05, +9.65]
Gemini	IL2	Correct-permuted	+63.85 [+35.65, +90.50]
Gemini	IL2	Correct-withheld	+83.25 [+59.45, +104.20]
Gemini	IL2	LLM-uniform	-29.75 [-51.25, -11.10]

independent starts. The small-model result still concerns a recovery-inclusive policy. The Gemini result shows that recovery is not necessary for the feedback/controller distinction in these conditions; it does not prove recovery irrelevant to every earlier result.

The two panels contain 320 and 160 new complete continuations, respectively. There are 1,440 logical LLM decisions: 360 LLM trajectories and 120 uniform-hit trajectories in total. The small-model scientific panel uses 1,260 physical requests; its separate technical pilot uses 108. Gemini uses 480. Accounted costs including conservative failures are \$0.68721517 and \$1.456788, respectively. No GPU was used. Source hashes, raw request/response records, parser/recovery provenance, acquired genes and own-history audits are retained with the experimental artifacts. The portable `confirmation_reproduction` directory reproduces the reported primary means and intervals from paired saved outcomes without network access.

A.4 Executable-batch preview development experiment

The development test fixes 40 historical mixed-policy states: Llama3.1/Qwen2.5 candidate generators, Carnevale/IL2, five starts per generator-source cell and the first/third adaptive rounds. The statistical unit is the parent start (20 total), not each of the 40 states or 240 calls. Each state supplies one saved LLM proposal and one uniform-hit proposal. Their centres and executed batches are fixed across views. Option A/B order is exactly balanced over states, shared across views and selectors, and hides generator identity.

Qwen2.5 and Gemini choose under three views: centre genes only; the correct 128-gene batch plus pre-test geometry and observed-hit-centre count; and the same preview payload assigned to the wrong candidate. All views include the same acquired true history and task. The correct preview exposes a computable action consequence, not unknown experimental outcomes. Deleting or changing candidate hit counts and truth records leaves every prompt unchanged. Gemini uses native JSON/low reasoning and Qwen uses a solution-only JSON-choice request. All 240 choices are valid on the first attempt, with no fallback. This manipulation bundles batch gene identities and geometric summaries; it does not identify their individual effects.

Correct preview changes 10/40 Qwen and 4/40 Gemini choices. Qwen’s preview-minus-centres contrast is +1.70 [-0.20,4.20] on IL2 and +0.25 [-0.30,1.10] on Carnevale; Gemini gives -0.45 [-1.35,0.00] and +0.30 [-0.20,1.10]. These small-source estimates are exploratory. Both selectors are hurt by deliberately misassigned previews, but that control alone cannot establish benefit from adding correct information. Neither model establishes a gain over centre-only information in this pilot. The geometry rule uses the same candidate pair without LLM calls, and neither selector establishes an advantage over it.

These results do not license a claim of a successful preview-based acquisition method, nor prove that every richer representation would fail. The candidate generators used historical contracts; the pilot does not regenerate their proposals under native modern interfaces. There is no full-campaign test or LLM selection between two uniform candidates here. The sample was not enlarged after

Table 11: Executable-preview development results, hits per next 128 tests. Mean two rounds within each start, then equally weight four generator–source cells. Exploratory 95% intervals resample five starts within each cell, 20,000 draws. These are not multiplicity-adjusted confirmation intervals or complete-campaign estimates.

Quantity	Qwen2.5	Gemini3.8Flash
Correct preview minus centres only	+0.98 [-0.08, +2.28]	-0.08 [-0.60, +0.40]
Correct minus misassigned preview	+6.08 [+3.90, +8.28]	+5.20 [+3.23, +7.35]
Correct preview minus uniform-hit	+1.00 [+0.30, +1.78]	+0.35 [+0.00, +0.78]
Correct preview minus geometry rule	-0.30 [-0.93, +0.15]	-0.95 [-1.98, -0.08]
Oracle minus correct-preview choice	+0.90 [+0.35, +1.55]	+1.55 [+0.70, +2.53]
Oracle minus centre-only choice	+1.88 [+0.85, +3.15]	+1.48 [+0.53, +2.60]

Qwen’s positive trend. This development test establishes sensitivity to incorrect previews, but does not establish an advantage from correct previews over centre-only information. Its model cost is \$0.496306.

Action/observation alignment and anticipated consequences are established ideas in agents [Yang et al., 2025b, Chae et al., 2025]. Our setting differs because the executed gene batch is exactly computable before testing, whereas its biological outcomes are unknown. Neither the general idea of previewing actions nor the absence of a significant gain in this small development set is claimed as a new universal principle.

B Recovery sensitivity, prior exposure and reproducibility

B.1 Incomplete proposals and finite-pool continuation bounds

The original 2,085 incomplete decisions span initializations, feedback arms, executors and budgets. Within the 720 B640 TRUE/NN LLM trajectories, 162 have at least one incomplete adaptive proposal; three shared initializations require recovery. Holding each realized shared initial batch and the uniform-hit trajectory fixed, we allow arbitrary continuation after the first incomplete adaptive proposal. If h hits have been observed among n tests, the eligible pool has H hits in N genes, and $r = 640 - n$ tests remain, the final LLM hit count is bounded by

$$h + \max\{0, r - [(N - H) - (n - h)]\} \leq Y_L \leq h + \min\{r, H - h\}.$$

For unaffected trajectories the observed outcome is retained. Averaging paired differences with the original equal-cell weights gives $[-17.70, +57.85]$ hits/640. This is a deterministic outer bound, not a confidence interval; arbitrary finite-pool completions need not be attainable by the NN executor. It preserves the observed prefix but allows all downstream choices to change. It does not replace recovered outcomes with invented model outputs, vary shared initialization, or identify a recovery-free model effect. Its inclusion of both signs shows that such an effect is unresolved without stronger assumptions or new collection.

Excluding Phi-4 gives $-9.54 [-10.75, -8.32]$ hits/640 over 600 starts/30 cells (supporting 95%, 20,000 stratified paired draws, seed 190919). Phi-4 is therefore not necessary for the observed deficit. Exclusion and bounds answer different questions: the former tests one-model sensitivity, the latter removes assumptions about continuation after incomplete proposals. The original complete-proposal subset also selects on realized performance of the interface and is not an unbiased estimate of the hypothetical model alone.

B.2 Publication status does not identify training exposure

Across the B640 sources other than CAR-T, LLM-minus-uniform-hit is $-11.93 [-13.30, -10.56]$ hits/640 (600 starts, 30 cells). CAR-T gives $-0.24 [-0.60, 0.12]$ (120 starts, six cells). Their descriptive difference is $-11.68 [-13.14, -10.27]$ hits/640. Intervals are supporting 95%, using within-cell resampling, not source-level population inference. The source-group difference uses independent cell-stratified draws (seed 190920); individual groups use seed 190919.

The larger deficit on published sources does not identify the effect of memorization. The groups differ in biology, hit prevalence, feature geometry and initialization. The CAR-T labels were publicly released even though the underlying experiment is described as unpublished, so they are not certified absent from pretraining. Unknown exposure can affect the shared LLM initialization and the subsequent opportunities available to both policies. A directionally conservative bias cannot be assumed for their paired difference. A future exposure test needs provenance or time-separated data, and a recall probe alone would not establish causal contamination.

B.3 Reproducibility checklist

Item	Retained material and boundary
Headline statistics	reproduction_core: 720 paired records, expected results, deterministic script and checksums; no paid calls.
Recovery and exposure checks	review_reproduction: per-start bound records, population counts and replay script; no raw feature values.
Full interfaces	Literal prompts, seeds, parsing, recovery and executor specifications in Appendix K.2.
Sources and outcomes	Source references, released endpoint definitions, eligible counts and retrieval identifiers in Appendix K.1.
Feature values	Not redistributed here; original feature input access and rights must be resolved separately.
Fresh model generation	Requires hosted-model access, specified serving routes and a budget; mutable aliases can prevent identical outputs.
Human comparison	Not collected. No synthetic expert responses are represented as human measurements.
Public availability	Source and saved-evidence packages are prepared locally, but are not embedded in this PDF; exact feature-input terms and anonymous hosting remain open. No public repository or independent full-pipeline reproduction is claimed.

B.4 Released AssayFormer: vocabulary-aware transfer

The released AssayFormer checkpoint, `model_last.pt`, provides a learned acquisition reference without new training or model API calls. We freeze all 120 original Carnevale starts (six model-specific initializers, twenty each), the original released hit authority, eligible pool and five 128-gene batches. The model ranks untested genes directly from the cached BioGRID 1900 screen description and cumulative binary observations. We reuse the authors' `RankerNet.encode` and `score_ids`, independently checking the latter against `score_all`. Ties follow lexical candidate order. Its training includes 1,349 historical screens; the Carnevale publication PMID 36002574 occurs only in the indicated test fold, while Shifrut occurs in train. This audit does not exclude all biological overlap or prior exposure in the description-embedding model.

The first adapter preserved the upstream unknown-token mapping for 301 genes outside the 24,217-token vocabulary. It selected 34,965 unknown genes across 61,440 adaptive slots, making its apparent deficit a representation failure rather than evidence against the learned method. We retain this run as a technical diagnostic. The subsequent evaluation restricted proposals to 16,835 representable genes and omitted unrepresentable initial observations from network context; all initial tests still count in utility and exclusion. This omits 255 of 15,360 initial observations from network context (0–6 per trajectory). The excluded 301 genes contain 15 of the 858 released hits: hit prevalence is 4.98% outside the vocabulary and 5.01% inside. This coverage check does not adjust for changes in adaptive histories or remove the action-space mismatch. The restricted evaluation uses existing embeddings and unchanged labels. Because its action-space restriction was chosen after inspecting the initial outputs, the result is exploratory.

The vocabulary-restricted model's mean yield is 48.99 hits/640, compared with 64.72 uniform-hit and 62.07 LLM. Paired learned-minus-uniform yield is -15.73 [$-17.68, -13.69$]; learned-minus-LLM is -13.08 [$-15.23, -10.91$]. The two comparisons use 20,000 within-initializer-cell paired bootstrap draws, seed 190921 and Bonferroni 97.5% intervals. Each uses the same 120 starts. The learned policy acquires zero out-of-vocabulary genes after initialization. It has additional historical

training labels, a different acquisition rule and a restricted proposal vocabulary. Those differences preclude an equal-information or controller-only superiority claim. We do not reproduce Assay-Loop’s native 1,000-gene benchmark or evaluate its LLM handoff. Original Schmidt prompt metadata do not resolve its multiple CRISPRa/i assay variants, so that source is not silently added to this transfer. Protocols, both sets of traces, input hashes and numerical replay are retained; checkpoint/data retrieval links are provided in the package.

B.5 Retuning the classical reference on separate LLM starts

The original classical reference was selected using random-gene development starts. A retrospective sensitivity instead uses saved LLM initializations from Qwen2.5 and Ministral on Carnevale and IL2. Four earlier starts per model/source (4001–4004) supply eight development initializations per source; all twenty balanced test starts (5001–5020) per model/source remain separate. These are previously studied fixed sources, not new biological validation. We freeze the original twenty-setting grid, run 320 development trajectories, select the highest mean hit yield separately per source with grid-order tie breaking, and evaluate the selected setting on all eighty test starts. Kernel, observations, feature pool, 128-gene batches and total 640-test budget are unchanged. No new model call is required.

Carnevale selects C17 (linear UCB, noise 10, $\beta = 2$); its development mean exceeds the runner-up by 1.00 hits/640. IL2 selects C00 (RBF GP, length scale 0.25, noise 0.01, $\beta = 1$); its development mean exceeds the runner-up by 0.50 hits/640.

Table 12: Classical-reference retuning sensitivity: mean hits/640 on twenty test starts per row. Original and retuned references, LLM and uniform-hit use exactly the same initial batch within each start. Settings are selected only from separate historical development starts.

Model initializer	Source	Original	Retuned	LLM	Uniform-hit
Qwen2.5	Carnevale	32.60	33.80	66.25	64.55
Qwen2.5	IL2	52.25	52.25	130.65	164.25
Ministral	Carnevale	35.45	35.50	55.65	63.10
Ministral	IL2	59.45	59.45	130.65	176.05

717

Across the four equally weighted cells, retuned-minus-original reference is $+0.31[-0.81, +1.47]$ hits/640; LLM-minus-retuned is $+50.55[+44.65, +56.40]$; uniform-hit-minus-retuned is $+71.74[+68.11, +75.04]$. These three retrospective comparisons use jointly adjusted 98.333% intervals (20,000 paired within-cell draws, seed 190923), conditional on the selected settings and the fixed screens. They do not include uncertainty over a new development sample. This checks historical two-model initialization matching; it does not certify an optimally tuned rival, cover all six original models, or replace a fresh development study. The untuned uniform-hit headline and original primary statistical families remain unchanged.

C Complete experimental definitions and study populations

C.1 Scientific task and model coverage

We measure retrospective single-gene hit acquisition. The 640-test (B640) panel uses Carnevale adenosine, Sanchez tau/down, Schmidt IL2/IFNG [Carnevale et al., 2022, Sanchez et al., 2021, Schmidt et al., 2022] and BDA’s released, unpublished CAR-T screen [Roohani et al., 2025]. Scharenberg choline recycling uses 160 tests (B160) [Scharenberg et al., 2023]. These tasks measure different biological responses: resistance to adenosine suppression, tau abundance, cytokine production, CAR-T performance and choline recycling. The seven readouts span five sources; Sanchez/Schmidt pairs are correlated.

We retain the released scores and hit labels. Genes must have Achilles features, numerical representations of how disrupting each gene affects cell viability across cell lines [BioDiscoveryAgent contributors, n.d., Broad Institute, n.d., Tsherniak et al., 2017, Meyers et al., 2017]. Table 13 gives the eligible populations, so a difference in recovered hits can be read against its source-specific

739 opportunity. For all seven original readouts, a hit means uppercase gene membership in the re-
740 leased `topmovers_SCREEN.npy` array, including CAR-T and Scharenberg. We do not reconstruct
741 those labels from raw counts or impose a common top-five-percent threshold; IL2 in particular dif-
742 fers from that rule. ORCS1107 instead retains the curated HIT flag, matching `Score.2 < 0.05` or
743 `Score.4 < 0.05`. These are distinct endpoint authorities, not a common biological threshold.

744 The manuscript source archive includes a saved-outcome replay:
745 `reproduction_core/reproduce.py` recomputes the three headline contrasts from 720 paired
746 records using Python and NumPy, without model calls or feature data. It checks summaries from
747 supplied outcomes, not the original experiments. The full handoff of responses and protocols
748 remains nonpublic, with feature redistribution unresolved. Internal replay therefore does not
749 establish independent reproduction of the complete pipeline (Appendices K.3 and K.1).

750 Five rounds acquire 128 genes each (B640), except Scharenberg’s native 32 (B160). We analyze
751 these budgets separately because they also represent different screens; their difference cannot iden-
tify a budget effect.

Table 13: Feature-eligible genes and released hits. Prevalence uses the eligible pool; the source labels are retained rather than reconstructed using a common quantile. Paired readouts from one study are not independent sources. ORCS1107 belongs to the separate external studies.

Readout	Eligible genes	Known hits	Hits (%)
Carnevale adenosine	17,136	858	5.01
Sanchez tau	17,204	859	4.99
Sanchez tau-down	17,204	883	5.13
Schmidt IL2	17,758	634	3.57
Schmidt IFNG	17,009	861	5.06
CAR-T	17,608	139	0.79
Scharenberg	1,013	48	4.74
External ORCS1107	17,883	3,538	19.78

752

753 The original balanced-panel checkpoints are Llama3.1-8B Instruct, Qwen2.5-7B Instruct, Gemma3-
754 4B IT, Ministral8B-2512, Qwen3-8B and Phi-4-14B [Grattafiori et al., 2024, Qwen et al., 2024,
755 Team et al., 2025, Mistral AI, 2025, Yang et al., 2025a, Abdin et al., 2024]. These general-purpose,
756 open-weight 4–14B models are evaluated on biomedical tasks. All six use complete histories, 384
757 output tokens and three attempts; Qwen3 thinking is disabled. Within each computational start, both
758 controllers receive the same LLM-selected initial batch. The initial condition is therefore matched
759 between policies; the estimate is conditional on this initialization rule, rather than averaged over all
760 possible starting rules. Batches differ across models, so model comparisons do not isolate capacity
761 [Bayley et al., 2026].

762 The original panel answers the primary question: does feedback help an LLM, and does that LLM
763 outperform an informed rival? Two sets of follow-up studies test the limits of this answer. External
764 checks use another screen; deployment and initialization checks change how the models are served
765 or how acquisition begins. Table 14 groups these three purposes. Each study is analyzed separately,
766 with its own model/screen subset. Gemini and DeepSeek enter the operating-condition checks,
767 bringing the total to eight language-model identities.

768 For the external checks, BioGRID ORCS1107, outside the BDA archive, measures primary-T-cell
769 proliferation [Shifrut et al., 2018, Stark et al., 2006]: 17,883 eligible genes/3,538 curated HIT labels
770 with native log fold change. The *external controller study* tests the original six models with correct
771 feedback and NN execution. The separate *external feedback study* tests Qwen2.5-7B/Phala and
772 Phi-4/DeepInfra, varying the outcomes shown to the controller. This separates an external check of
773 controller performance from a check of how feedback changes that performance. We do not pool the
774 two populations. Prior exposure and shared biology/laboratories limit generalization to independent
775 sources.

Table 14: Claim-specific coverage. S/R: sources/readouts; policies: arms per start/recipe. A recipe specifies a model with its serving and prompting settings. Starts are computational, not biological replicates; their counting unit is shown in each row. Row groups distinguish the original panel, external checks and operating-condition checks. Budgets count genes.

Study	Models/recipes	S/R	Starts	Policies	Budget
Original B640 core	6/6	4/6	720 total	14	640
Scharenberg boundary	6/6	1/1	120 total	14	160
External controller study	6/6	1/1	120 total	4	640
External feedback study	2/2	1/1	40 total	6	640
Deployment study	3/4	2/2	40 per recipe	8	640
Qwen initializer discovery	1/1	2/2	40 total	4	640
Qwen initializer confirmation	1/1	2/2	80 total	5	640
Initializer-controller study	2/2	2/2	80 per model	10	640

Original arms: Figure 1. Deployment recipes share uniform-hit arms; initializer confirmation adds a later matched control. The initializer-controller study shares four uniform policies between models (1,280 unique trajectories).

776 C.2 Controllers and experimental conditions

777 Each round has three steps: propose genes, expand them into a batch and reveal recorded outcomes.
778 The controller receives the tested genes and the feedback allowed by condition a . We parse its
779 response for legal gene symbols. If too few remain after all attempts, a deterministic hash-based
780 rule fills the gaps without using outcomes; we call this *recovery*. The proposal centres C_r therefore
781 include both controller and recovery outputs. Formally,

$$C_r = p(F_a(H_r)), \quad A_r = E(C_r, H_r), \quad H_{r+1} = H_r \cup \{(g, y_g) : g \in A_r\}.$$

782 Here H_r is the acquired history, F_a renders the allowed feedback, p includes the controller, parsing
783 and recovery, E is the executor, A_r is the new batch and y_g is a recorded score/hit pair. For example,
784 after an initial 128-gene batch, a controller reads the available history and proposes five centres.
785 Parsing and recovery produce a legal proposal; NN then selects untested neighbours until the next
786 128-gene batch is full. The screen supplies those genes’ recorded outcomes. A tested hit may guide
787 this search without being tested again. This illustrative round repeats on each policy’s own history,
788 so initially matched policies can later see different outcomes. The architecture follows Gupta et al.
789 [2025]; BDA search/critic tools are absent.

790 In the original panel, the system prompt states the screen task, measurement and five-round budget.
791 The adaptive user message includes the complete acquired history as gene-name/score tables headed
792 [HITS] and [OTHER RESULTS]; NO instead lists tested names without outcomes. Its instruction
793 includes: “Update your priors appropriately and choose genes that gave you hits. Also, be sure to
794 explore by including some genes that could give hits.”

795 It requests Reflection, Research Plan and Solution fields, with five HGNC symbols in Solution. The
796 parser reads comma/newline-separated entries after Solution, normalizes the first token to uppercase,
797 and retains unique eligible symbols across up to three attempts. Recovery fills missing centres in
798 a fixed SHA-256 order keyed by model, screen, start and round, independent of feedback arm and
799 outcomes. In centre order, NN fills quotas of 26/26/26/25/25 genes (or 7/7/6/6/6 at B160), skipping
800 tested or already selected genes. Full literal prompts remain in Appendix K.2.

801 TRUE preserves each gene’s score/hit pair. WITHIN reassigns intact score/hit pairs across genes
802 within each acquired batch, preserving tested identities, batch membership, the outcome multiset
803 and hit prevalence. NO withholds outcomes but retains tested identities for exclusion. Uniform-hit
804 samples five displayed hits; first-hit takes the first five in recorded acquisition order: earlier batches
805 first, then the recorded gene order within each batch. It does not rank hits by score or gene name.
806 Under NO the same rules select from tested identities without hit filtering. A displayed hit is a gene
807 labeled as a hit in the supplied feedback; after permutation, it need not be a true hit for that gene.
808 Both rules use the same recovery. A centre may already have been tested: NN selects its untested
809 neighbours. The original balanced panel starts from an LLM-selected batch; the deployment study
810 starts from uniformly sampled genes.

811 NN expands centres through cosine-nearest untested genes in normalized Achilles space. RANDOM
812 EXPANSION retains exactly the centres NN would include at that arm’s current state and uniformly
813 fills remaining legal non-centre slots. For both correct-feedback controllers, this changes batch
814 expansion while preserving centre inclusion at the current state. This changes the neighbours, not
815 simply the entire policy to random acquisition. Subsequent histories may diverge. Original prompts
816 remain NN-oriented; deployment-study prompts explain the actual executor.

817 Random acquisition samples untested genes uniformly without replacement; Coreset greedily max-
818 imizes minimum cosine distance to tested and newly selected genes [Sener and Savarese, 2018].
819 Both use shared initialization/features without outcomes. Uniform-hit, first-hit, Gaussian-process
820 upper confidence bound (GP-UCB) and linear-UCB complete six baseline families.

821 We tune GP-UCB and linear-UCB rather than compare arbitrary settings: 20 settings on four sepa-
822 rate development starts/screen (480 trajectories), selecting by yield with fixed tie breaking. To select
823 a batch before its outcomes arrive, these rules provisionally substitute predicted outcomes (posterior-
824 mean “fantasies”). Full refits check this approximation [Li et al., 2010, Srinivas et al., 2010, Snoek
825 et al., 2012]. CAR-T/external tests transfer the best cross-screen normalized-yield setting; test fit-
826 ting uses acquired outcomes only, without retuning. The selected reference is GP-UCB except on
827 Sanchez tau, where it is linear-UCB. The chosen setting survives 0/4 to 4/4 leave-one-development-
828 start-out checks across screens; its development lead over the runner-up is only 0.25–1.50 hits. Test
829 intervals condition on this selected setting rather than incorporating selection uncertainty. Develop-
830 ment used random-gene starts whereas the original test panel uses LLM-selected starts, so this is a
831 specified reference under a distribution change, not an optimally tuned upper bound (Appendix I).

832 The operating-condition checks answer three different questions. The *deployment study* changes
833 serving and prompting while keeping random-gene starts: does the controller comparison change?
834 The *Qwen initializer study* tests alternative starting batches for Qwen. The *initializer-controller*
835 *study* crosses three starting rules with two continuation controllers for Gemini and DeepSeek: does
836 the better controller depend on how acquisition begins? These studies have separate populations and
837 statistical comparisons, as described below.

838 Table 15 specifies the serving and response contracts. All recipes expose complete acquired histories
839 and allow at most three content attempts to obtain five legal centres, followed by outcome-blind
840 recovery. Each content attempt permits up to four transport attempts in these hosted campaigns.
841 JSON recipes explain the actual executor; the legacy Qwen recipe retains the original NN-oriented
842 instructions.

843 The deployment study starts every arm from random genes. The Qwen initializer studies use the
844 JSON Qwen recipe; the initializer-controller study retains the listed Gemini/DeepSeek recipes while
845 varying the starting rule. Hosted aliases do not attest immutable weights. Qwen’s token, format
846 and instruction changes are bundled, so their comparison does not isolate any one ingredient. Full
847 payloads and seed contracts are in Appendix D.4.

Table 15: Hosted serving contracts. Limits are requested output tokens per attempt; temperature is a sampling setting. These rows specify deployments, not isolated model-size comparisons.

Recipe	Provider	Tokens	Temp.	Response / reasoning
Gemini3.8Flash	Google AI Studio	4,096	1.0	JSON; low reasoning
DeepSeekV4	OpenInference fp8	2,048	0.6	JSON; reasoning off
Flash0731				
Qwen2.5-7B legacy	Phala	384	0.2	Original three fields
Qwen2.5-7B JSON	Phala	1,536	0.2	JSON centres only

848 The deployment study pools Gemini and DeepSeek on Carnevale and Schmidt IL2 for five primary
849 controller, feedback and executor comparisons. Qwen’s serving comparisons remain separate. The
850 initial batch is random in every arm, so this study does not estimate an initializer effect.

851 The Qwen initializer study tests LLM-selected starting batches in a discovery panel and a larger
852 confirmation panel with a matched control (Table 14). Its interaction and matched-control intervals
853 are in Appendix D.4.

The *initializer-controller study* asks whether the best controller depends on the starting batch. Gemini and DeepSeek each use forty fresh starts/source on Carnevale and Schmidt IL2, the same two previously used sources. We cross LLM/uniform-hit continuation with three initializers: random genes; random centres expanded by NN; and LLM centres chosen without outcomes, also expanded by NN. Continuation uses correct feedback. Random-gene starts additionally cross WITHIN/NO. The initial 128 tests count toward 640. We test six planned contrasts for each model as one twelve-test family. Table 3 reports the adjusted intervals; Section C.3 describes their calculation. Response settings, decision counts and full protocols are in Appendix D.

C.3 Statistical definitions and diagnostic analyses

Appendix M gives the complete contrast definitions, multiplicity families, resampling units, prediction folds, annotation formulas and retrospective graph controls. The study-specific sections report their populations and numerical results. The intervals condition on the fixed biological sources; repeated computational starts do not create biological replications. Earlier failed horizon predictions and later retrospective checks remain distinguished in Appendix J.

Table 16: Paired hits/640 with Bonferroni 98.75% intervals: separate four-contrast families for original B640 (720 starts,36 cells) and external feedback (40 starts,two models). Δ : LLM-minus-uniform-hit; Γ : difference in assignment benefits; Ω_L/Ω_U : NN expansion benefit for LLM/uniform-hit; $B_{\text{assign}}/B_{\text{avail}}$: LLM assignment/availability benefit. Sources are fixed.

Original balanced panel		External feedback cross	
Contrast	Hits / 640	Contrast	Hits / 640
Δ	−9.98 [−11.47,−8.56]	Δ	−33.55 [−56.28,−13.42]
Γ	−12.06 [−14.40,−9.72]	Γ	−71.83 [−108.68,−36.82]
Ω_L	32.27 [30.59,33.86]	B_{assign}	10.98 [−15.72,35.15]
Ω_U	43.41 [42.33,44.42]	B_{avail}	58.85 [32.20,82.60]

D Initializer-controller interactions

The fresh experiment uses 80 computational starts on two previously exposed sources, not a new biological population. Main Table 3 contains the complete twelve-contrast primary family. The initializer interaction compares the LLM-minus-uniform difference after LLM versus random-centre NN initialization; it does not compare significance labels. Feedback interventions are crossed only after random-gene initialization. Consequently, outcome benefit in that condition cannot establish feedback–advantage dissociation after LLM initialization.

D.1 Fixed protocol, sampling and failures

All three initializers acquire 128 eligible genes; initialization is included in the five-round 640-gene budget. The random-centre initializer samples five centres without outcomes and uses exactly the same nearest-neighbour expansion as the LLM initializer. Both continuation controllers receive identical information contracts and follow their own complete histories. Four non-LLM trajectories are shared between models; there are 1,280 unique trajectories rather than 1,600 independent ones. The source-stratified bootstrap resamples paired starts jointly across conditions and models. Each of twelve primary intervals has nominal 99.583% coverage (Bonferroni family 95%); bootstrap coverage is approximate. No new significance-driven sampling or equivalence margin is introduced. Forty starts/source was fixed before collection using the prior experiment’s cost and variance as imperfect planning proxies; it was not selected after looking at new yields.

The new recipes retain Google AI Studio and OpenInference/fp8 routing, three proposal attempts, JSON output and outcome-blind recovery. Response records return the requested Gemini3.8Flash and DeepSeekV4Flash0731 identities; these aliases do not attest immutable hosted weights. Gemini completed 1,680/1,680 legal five-centre decisions in 1,684 requests; DeepSeek completed 1,678/1,680 in 1,824 requests. There were no transport errors. Two DeepSeek decisions used recovery, retained in end-to-end performance. Four synthetic-only requests cost \$0.004874216; 3,508 main requests cost \$5.439075107, total \$5.443949323. Local execution is excluded from API time.

Every stored prompt, transformation, parser/recovery action, NN acquisition and truth lookup was replayed, including 5,120 adaptive decisions and 320 unique initializations. All 3,512 physical requests reconcile to the cost ledger. Saved-outcome reproduction is separate from paid stochastic regeneration.

D.2 Source-specific effects

The following source intervals are supporting 95%, not an additional multiplicity-controlled primary family. Sources receive equal weight in the main; variation across computational starts does not identify variability across independent biological studies.

Table 17: Gemini3.8Flash source-specific paired differences in hits/640; supporting 95% intervals.

Contrast	IL2	Carnevale adenosine
Controller after random genes	2.90 [-1.43,7.58]	3.95 [0.70,7.35]
Outcome availability (random genes)	127.12 [120.08,133.30]	25.15 [22.07,28.10]
Correct assignment (random genes)	73.05 [59.02,86.83]	9.75 [5.47,14.18]
Controller after LLM initialization	-52.00 [-66.95,-37.30]	6.28 [4.08,8.68]
LLM minus random-centre initializer	-30.07 [-35.75,-24.52]	-0.78 [-5.72,4.50]
Initializer $\times$ controller interaction	-49.62 [-65.15,-34.35]	0.55 [-3.10,4.03]

Table 18: DeepSeekV4 source-specific paired differences in hits/640; supporting 95% intervals.

Contrast	IL2	Carnevale adenosine
Controller after random genes	6.30 [1.25,11.70]	3.92 [-0.50,8.28]
Outcome availability (random genes)	110.47 [104.15,116.70]	27.60 [22.93,31.60]
Correct assignment (random genes)	53.55 [39.50,68.20]	7.50 [2.80,12.30]
Controller after LLM initialization	-25.43 [-39.65,-12.35]	-7.10 [-11.43,-2.85]
LLM minus random-centre initializer	-14.68 [-19.32,-9.90]	-7.05 [-11.65,-1.95]
Initializer $\times$ controller interaction	-24.38 [-39.65,-9.78]	-13.12 [-17.50,-8.85]

D.3 Absolute operating points and boundaries

Table 19 lists all cells. Uniform policies after random initialization are shared across model comparisons. LLM-initializer calls belong to either deployable continuation policy but are counted only once in actual expenditure. Correct-feedback benefits after LLM initialization are untested; a continued-controller loss is insufficient to infer that feedback is unhelpful.

Model	Initializer	Controller/feedback	Mean hits
Shared	Random genes	Uniform/TRUE	109.31
Shared	Random genes	Uniform/WITHIN	75.80
Shared	Random genes	Uniform/NO	65.11
Shared	Random centres	Uniform/TRUE	118.34
Gemini3.8Flash	Random genes	LLM/TRUE	112.74
Gemini3.8Flash	Random genes	LLM/WITHIN	71.34
Gemini3.8Flash	Random genes	LLM/NO	36.60
Gemini3.8Flash	Random centres	LLM/TRUE	120.01
Gemini3.8Flash	LLM prior	LLM/TRUE	80.05
Gemini3.8Flash	LLM prior	Uniform/TRUE	102.91
DeepSeekV4	Random genes	LLM/TRUE	114.42
DeepSeekV4	Random genes	LLM/WITHIN	83.90
DeepSeekV4	Random genes	LLM/NO	45.39
DeepSeekV4	Random centres	LLM/TRUE	120.83
DeepSeekV4	LLM prior	LLM/TRUE	91.21
DeepSeekV4	LLM prior	Uniform/TRUE	107.47

Table 19: All fresh modern cell means (hits/640): 80 starts/cell; shared controls listed once.

Per-trajectory API use. For random-initialized TRUE/NN trajectories, Gemini/DeepSeek/uniform-hit cost \$0.01491/\$0.00031/\$0 per trajectory. Median summed API-attempt times are 7.51/9.22/0 seconds, including repairs but excluding local execution and scheduling; zero API cost is not zero system cost.

D.4 Earlier initialization confirmation and its boundary

The earlier Qwen2.5-7B study used Carnevale and IL2, with 40 fresh starts per source (16001–16040) after an exploratory comparison. The LLM proposed five centres before outcomes were available; NN expanded them into 128 initial tests. Uniform-hit and Qwen then shared that initial batch and made four further 128-test acquisitions. The rival initializer sampled 128 individual genes uniformly. Confirmation collected both initializers together; the earlier discovery comparison had also differed in serving epoch.

Qwen used the Phala route, temperature 0.2, a JSON centres array and 1,536 output tokens, with at most three content attempts and outcome-blind recovery. All 320 confirmation trajectories completed; 78/80 initial and 637/640 adaptive LLM decisions were complete before recovery. The primary initializer–controller interaction was -4.50 $[-9.86, 0.79]$ hits/640 (95% paired interval), establishing neither an interaction nor equivalence. The prior-only initial-yield difference was $+3.075$ $[2.0875, 4.0501]$ hits (supporting 99%). Starts and sample size were fixed before confirmation; collection was not enlarged after this null.

A post-confirmation matched-initializer check kept those 80 starts and uniform-hit continuation, but selected five outcome-blind random centres followed by NN initialization. Direct-random, random-centre NN and LLM-centre NN starts produced 110.8625, 117.7500 and 115.3375 final hits. LLM minus matched-random initialization was -2.4125 $[-5.5375, 0.7377]$; matched-random minus direct-random was $+6.8875$ $[3.1873, 10.5375]$ (supporting 97.5% paired intervals). The earlier LLM-versus-direct-random comparison therefore does not establish LLM-specific prior value: it also changes NN initialization. This added arm was not a full three-initializer-by-two-controller factorial. The separate modern-model factorial is reported in Appendix D.

The full version retains the preceding 800-trajectory response-contract cohort, its failed technical pilots and excluded invalid-seed cohort, all initialization development details, raw-response accounting and original five-contrast inference. These are separate experiments, not extra observations in the modern initializer cross or the full-campaign response checks.

E Controller-selection diagnostics

These extensions reuse the already exposed biological screens. New own-history model calls are fresh generations on 140 reused initial batches, not new independent biological sources. All initial source labels, score tables, hit indicators and feature pools are unchanged.

E.1 Direct centres, neighbourhoods and prediction

All 13,440 TRUE adaptive decisions (840 starts, two controllers, two executors, four adaptive rounds) were decomposed into directly acquired centre hits and other hits. Initial batches cancel in paired contrasts. The random-expansion expected background uses its remaining non-centre population, so outcome-informed accounting is distinguished from usable predictions. The B640 background residual difference is +0.15 [-0.30,0.62]. Component subtraction is not an intervened campaign.

At the first shared adaptive state, the equal-budget exclusion executor removes all proposed centres from eligibility before reapplying the original quotas, ordering and deduplication. Its LLM-minus-uniform B640 difference is -2.42 [-2.97,-1.88], versus -2.22 [-2.78,-1.69] under the original NN executor.

The fixed ridge predictor uses differences in centre reuse, actor completeness, non-centre distance, log neighbour rank, inter-centre cosine similarity and observed eight-neighbour hit support. Standardization and fitting use training source groups only, with fixed regularization 10 and equal source weights. Schmidt readouts and Sanchez directions stay grouped. All test outcomes are scoring-only; mutating them leaves features unchanged. This retrospective cross-fit is not untouched-source validation; its bootstrap conditions on fitted fold models.

The geometry repair prefers the lower mean distance among non-centre acquired neighbours. It ties in favor of the non-LLM candidate. Uniform-two supplies two independently sampled observed-hit centre sets under the same criterion. All four policies retain separate evolving histories, identical initial batches, original 384-token prompts, temperature 0.2, three content attempts, fixed routes and outcome-blind shortage recovery. The development-selected classical results are reused only after exact initial-state and source-hash verification.

Table 20: Geometry selection minus its rivals, by screen. Supporting 95% intervals; B640 pooled first two contrasts use 97.5% in the main paper. Each row contains 20 paired starts across two checkpoints.

Screen	Uniform-hit	Uniform-two	Selected classical
Carnevale	+3.45[+1.75, +5.35]	+1.45[-0.30, +3.25]	+37.00[+33.10, +40.35]
Sanchez	+5.45[+2.15, +8.95]	+2.85[-1.65, +7.45]	+19.75[+15.85, +23.45]
Sanchez-down	+2.95[+1.00, +5.40]	+1.00[-0.80, +3.00]	-1.00[-5.10, +3.05]
Scharenberg	+0.25[-0.15, +0.70]	+0.15[-0.45, +0.75]	-3.55[-4.85, -1.90]
CAR-T	-0.15[-0.65, +0.45]	-0.20[-0.75, +0.40]	+0.50[-0.55, +1.50]
IL2	+2.30[+0.25, +4.80]	-0.80[-2.70, +1.00]	+118.05[+112.25, +123.65]
IFNG	+2.70[+0.60, +5.05]	-4.95[-7.85, -1.85]	+47.10[+41.15, +53.55]

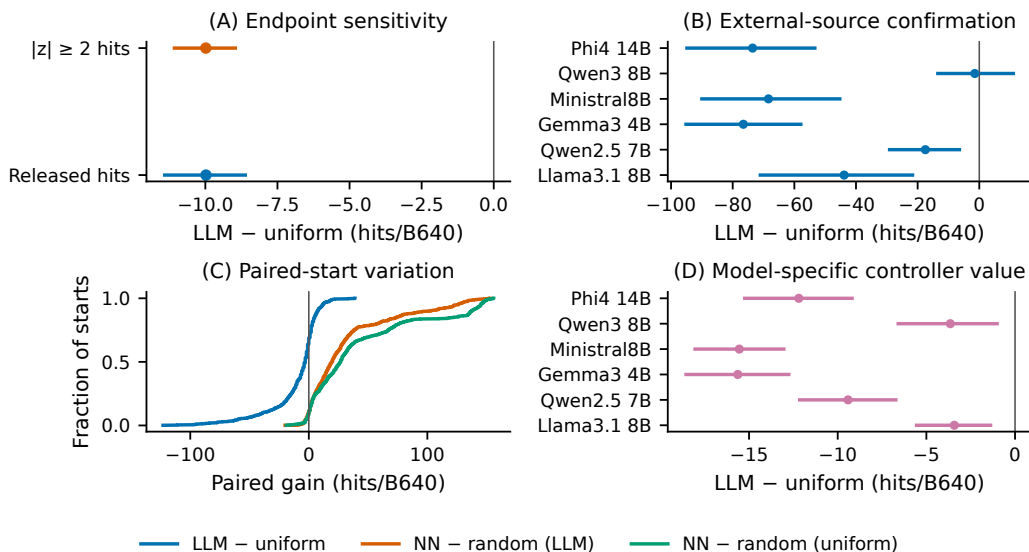


Figure 7: Retained endpoint sensitivity, external-source model contrasts, paired-start distributions and model-specific controller value. The original adjusted and supporting intervals remain unchanged. ECDFs retain all 720 balanced B640 starts.

963 E.2 Predicting controller advantage from the observed history

964 An informed hit-driven controller may have less to exploit when the acquired history contains few
 965 useful neighbourhoods. We test whether observable neighbourhood opportunity predicts where the
 966 LLM adds next-batch yield. Unlike the earlier proposal-difference predictor above, these features are
 967 available before either controller proposes centres. This is an exploratory challenge on the original
 968 720 B640 first-adaptive states, not new biological validation or a test across the modern study's
 969 initializers. Each pair shares its 128 acquired genes and observed outcomes; its target is LLM-
 970 minus-uniform-hit yield on the next 128 tests. Repeated initializations give 667 distinct acquired-
 971 gene sets within model-readout groups; starts remain computational repetitions. All 1,440 retained
 972 next batches and their released-hit counts were checked against the original records.

973 The first feature is $\log(1 + h)$, where h is the number of acquired hits. For each hit we find its
 974 26 nearest untested neighbours in the same Achilles representation; 26 is the largest centre quota,
 975 fixed before fitting. Three added features are the union size divided by the remaining eligible pool,
 976 overlap $1 - |\text{union}|/(26h)$, and the fraction of each hit's original 26 nearest other genes already
 977 acquired. For $h = 0$ all three equal zero. Coverage measures how much of the untested pool
 978 is reached by these neighbourhoods; overlap measures repeated coverage, and depletion measures
 979 how much of the original neighbourhood has already been tested. None uses the hidden hit rate of
 980 untested neighbours. Feature construction uses acquired labels only; altered future outcomes and
 981 LLM proposals leave it unchanged. Cached ranks were checked against direct recomputation on
 982 each readout.

983 We compare a training-mean constant, hit-count-only ridge regression and ridge with all four fea-
 984 tures. Each fold excludes one originating source, keeping Schmidt and Sanchez readouts together.
 985 Training sources receive equal weight; standardization, intercept and coefficients use training folds
 986 only. Regularization is fixed at 10 without a parameter search. Errors below are averaged equally
 987 across the four held-out sources; all sources had already informed development. Numerical out-
 988 comes use one retained proposal per state. Thus these errors are neither untouched-source confirma-
 989 tion nor a comparison with an optimal numerical controller.

Table 21: Source-excluded next-batch prediction. Mean squared error is in squared hits per 128 tests; lower is better. The LLM / uniform column gives absolute next-batch mean yields. The macro row weights sources equally, not individual states.

Source	Starts	LLM / uniform	Constant	Hit count	Opportunity
CAR-T	120	1.40 / 1.12	9.23	13.43	13.81
Carnevale	120	13.29 / 13.81	28.28	33.32	34.79
Sanchez	240	8.99 / 8.40	27.49	33.01	33.15
Schmidt	240	21.62 / 28.74	213.89	210.55	209.10
Macro	720	—	69.72	72.58	72.71

Opportunity improves squared error over hit count on only one source. The macro improvement (hit-only error minus opportunity error) is -0.14 $[-1.13, 0.87]$. Choosing the LLM only when the predicted difference is positive gains $+0.06$ $[-0.07, 0.19]$ hits/128 over the hit-only rule, but its mean gain over always using uniform-hit is -0.16 . These exploratory 95% intervals resample 20,000 paired parent starts within model/source, keeping correlated readouts together. They condition on fitted folds and the four selected sources; they omit training-fit and new-source variability. Adding identical deployment indicators to the three specifications gives macro errors 69.90, 72.63 and 73.15, preserving the lack of improvement.

The pre-analysis continuation rule required error improvement on at least three sources and an average selection increment of 0.5 hits/128 over hit count. That increment was a planning threshold, not a biological utility claim. Neither requirement was met, so no predictor-driven campaign or conditional modern-initializer stress test was launched. This result limits the tested state features and linear predictor, not every account of hit scarcity or source reversals. Full features, fold predictions, source metrics and executable statistical checks are in `state_opportunity_reproduction`; raw-feature regeneration requires the separately identified Achilles inputs.

F Balanced feedback-executor experiment

The balanced experiment uses start identifiers 5001–5020 and a distinct generation namespace. The complete frozen protocol, request payloads, outcomes, deterministic recovery and cost ledger are retained in the local experimental archive; the submission PDF contains their protocol and numerical summaries. Each of the 840 starts includes nine NN controller-feedback trajectories, two TRUE-feedback random-expansion trajectories and one selected classical trajectory. Each has five rounds, giving 10,080 core test trajectories. An additional 1,680 outcome-blind random and coreset trajectories use the same starts, for 11,760 total trajectories. The shared initial LLM decision plus sixteen adaptive LLM decisions per start gives 14,280 logical decisions; copied initial batches do not count as extra inference. All 50,400 trajectory rounds are checked for budget, uniqueness, exclusion and released outcomes. Deterministic executor replay checks the final adaptive round of all eleven nonclassical policies in each of 42 first-start cells (462 rounds). Independent no-feature replay reconstructs every cumulative endpoint and the four primary intervals from released truth. Response-to-decision replay also checks all 37,800 unique current controller decisions: regenerated feedback packets and request payloads, response parsing, simple-control choices, deterministic recovery, slot structure and released truth. It reconstructs all 6,720 random-expansion actions conditional on recorded NN centre inclusion; full NN geometry still requires the feature matrix.

F.1 Serving contract and cohort amendments

A six-block pilot qualified legal generation and transport without admission by biological yield. The retained cohort has 25,339 physical attempts. The initial DeepInfra Llama route was replaced as an entire partial model stratum by Groq after technical qualification; its 36 retired attempts remain archived and conservatively charged at \$0.002069, outside the scientific estimates. Hosted aliases do not establish identical hidden weight revisions across routes.

Rate-limit rejections triggered identical-payload transport retries and bounded adaptive concurrency, with a final lifetime bound of 48 physical attempts per content attempt. Transport rejection does not consume one of the three content attempts. Request payloads, prompts, seeds, sample membership, endpoints and statistical families were unchanged by transport handling. All failures, partial deci-

sions and retired cohorts are retained. Service-health probes and three local scheduler-test events are excluded from scientific observations; the health probe costs \$0.0002. Biological outcomes did not select routes, model capacity or continuation.

F.2 Source-conditioned numerical boundary

For uniform-hit NN minus the selected classical rule, mean released-hit differences and supporting 95% intervals are: Carnevale22_Adenosine: 31.89 [30.33,33.37]; IFNG: 47.88 [45.54,50.14]; IL2: 115.59 [112.24,118.75]; Sanchez21: 19.39 [17.40,21.26]; Sanchez21_down: -2.50 [-4.30,-0.66]; Scharenberg22: -3.23 [-3.93,-2.47]; Steinhart_crispra_GD2_D22: 0.67 [0.28,1.05]. Scharenberg uses B160; other screens use B640. The random and coresets references were specified before complete-panel outcome inspection, with no tuning or alteration of the four primary tests. Their additional 8,400 round-endpoint vectors are independently reconstructed. The machine-readable supplement retains every model-screen interval, endpoint, trajectory, leave-model/family/source estimate and complete-actor selection count. Figures in the main paper carry the central source and robustness evidence; additional numerical cells do not add replications.

F.3 Dependence and replay

Intervals condition on fixed sources and implementations. Starts share gene universes, candidate features and released measurements; paired policies and repeated rounds are nested observations. The two Sanchez directions and two Schmidt readouts remain correlated. Supporting source-equal weighting averages the fixed B640 source groups, then resamples starts within cells; it is not a biological-study bootstrap. CAR-T uses a rule transferred from the prior development grid and has no source-specific tuning. Whole-screen z-score moments are used only for evaluation. No retrospective model omission, outcome-dependent top-up or equivalence margin was introduced.

Interrupted-request retention. Collection paused without dropping completed or rejected requests. Continuation resumed only unserved decisions, retaining saved partial responses and rejected attempts. Raw request statuses remain available; some earlier error-body metadata were not captured and cannot be reconstructed. This affects transport diagnostics, not the released-outcome authority.

G External curated-screen protocol and source boundary

The external confirmation was frozen before completed balanced-panel biological estimates were inspected. Start identifiers 7001–7020, a separate generation namespace, and the same six deployed checkpoints yield 120 shared-initialized starts and 480 trajectories. Its 600 logical LLM decisions are separate from the balanced 14,280. All 1,560 unique controller actions undergo raw-prompt, parser, recovery and full nearest-neighbour feature replay; 2,400 trajectory rounds are checked against released truth. The external primary is one pooled LLM-minus-uniform-hit contrast, using 20,000 model-stratified paired bootstrap replicates, seed 80909, 95% intervals. Per-model, first-hit and selected-GP comparisons are supporting. No model, start or endpoint is omitted by outcome.

The operational source is the BioGRID ORCS 2.0.18 screen 1107 matrix attributed to Shifrut et al. [2018]. We preserve all 19,100 downloadable unique official gene symbols and its 3,787 HIT flags. Every flag matches the release’s stated $\text{Score.2} < 0.05$ or $\text{Score.4} < 0.05$ rule; Score.5 supplies native log fold change. The index’s 19,114 original-row count exceeds the actual download by 14, which are not imputed. Exact uppercase feature matching retains 17,883 genes/3,538 hits. The linked original workbook contains duplicated/date-corrupted identifiers and its numerical fields do not reproduce the curated hit count; it is not silently substituted. This limits interpretation to the stated curated-release endpoint. The raw matrix, source index, conversion, hashes and source checks are archived; private workbook metadata are excluded. The new native scores are neither identical nor affine duplicates of the seven BDA-release score vectors. That check cannot establish independent donors or laboratories. Related T-cell biology and contributing laboratories remain a transport limitation.

The external GP uses the earlier development-selected transfer setting C00, without tuning on this source. Its adaptive fits observe only acquired hit outcomes. The external cohort retained all com-

1082 pleted starts when its transport wrapper adopted the same bounded retry and admission policy. Sci-
 1083 entific settings and the separate cost ceiling did not change. Uniform-hit minus this GP is 31.31
 1084 [25.13,37.29] hits/B640 (supporting 95%). External API expenditure is \$0.31. These campaigns
 1085 reveal outcomes from a released dataset rather than conducting new wet-lab experiments.

1086 G.1 Full cell, smaller-budget trajectory and operating surfaces

1087 The following figures show per-cell contrasts, trajectories at each native budget and inference costs
 for the completed balanced panel.

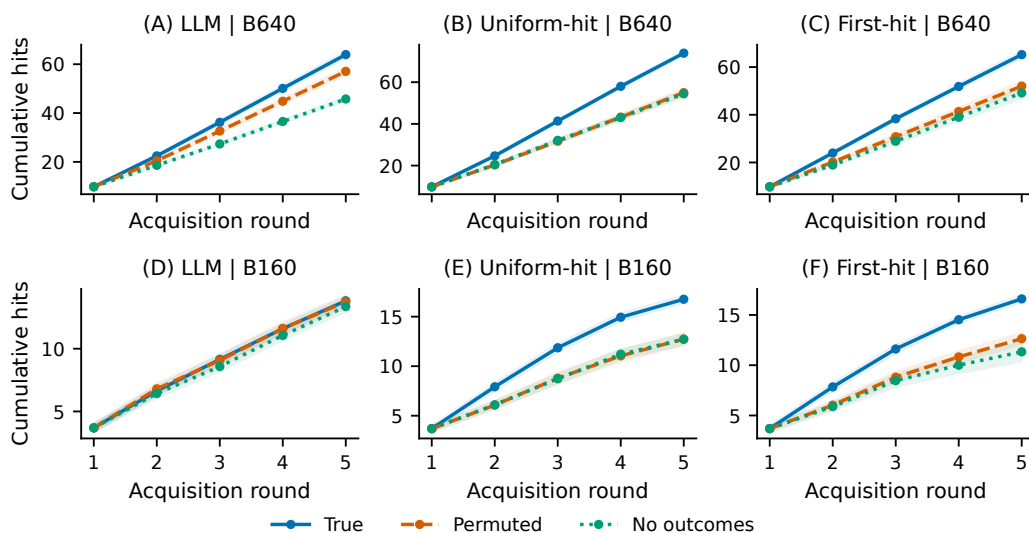


Figure 8: Full feedback trajectories at B640 (720 starts) and B160 (120 starts); supporting 95% intervals.

1088

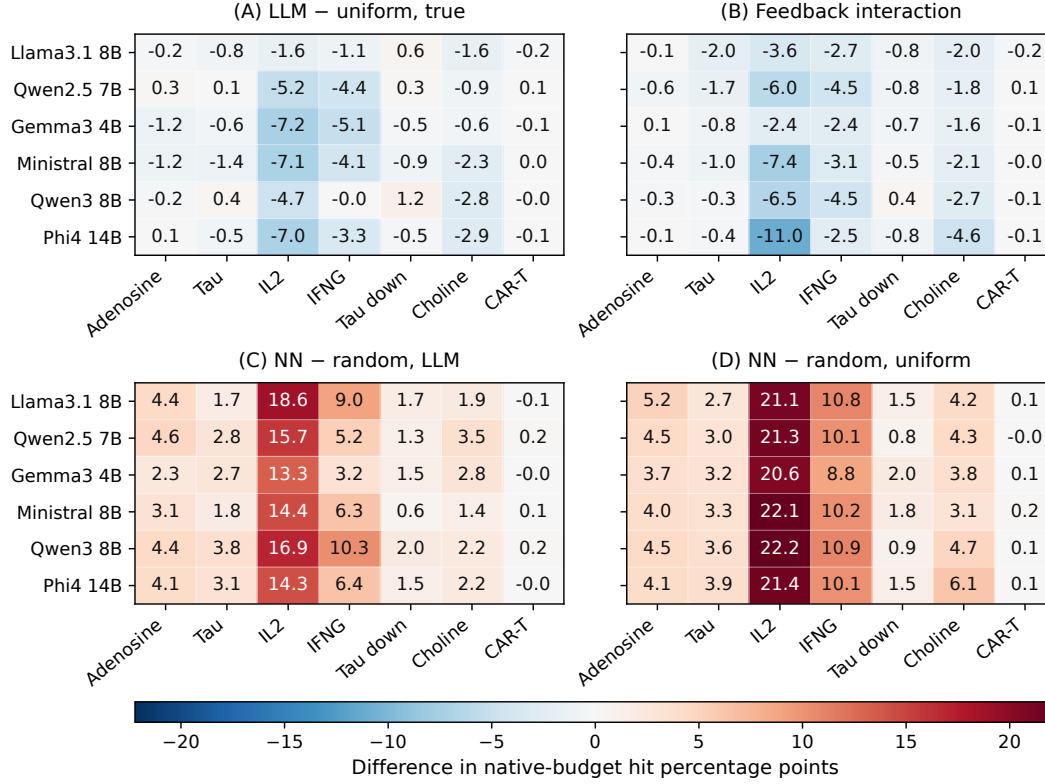


Figure 9: Every balanced model–screen cell for the four primary contrasts. Values are descriptive mean differences in native-budget hit percentage points, not significance annotations. All four panels share one zero-centred colour scale; each printed cell is the original descriptive mean.

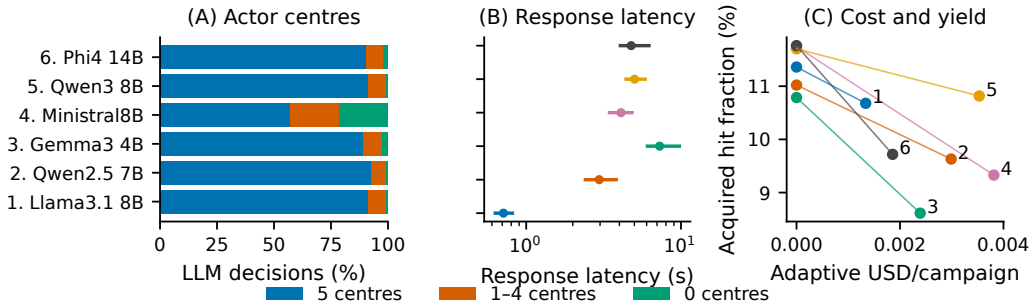


Figure 10: Original whole-campaign actor completion, successful-response latency (median/IQR), and adaptive inference cost versus yield. Initial inference, local computation and development are excluded from the right-panel cost axis. Each right-panel segment connects a model-stratum uniform-hit reference at zero adaptive API cost to the paired LLM policy at its recorded mean cost. Segments identify the two observed configurations, not an interpolated cost–yield frontier; labels 1–6 follow the model order in the left panel.

H Matched interface controls

The naming and output-budget protocols were frozen before their paid calls. Both reuse the 840 balanced initial states and evaluate one next batch. The 720 B128 states form the primary population and the 120 B32 states a separate supporting population. Each protocol freeze defines its own primary testing family; these intervals do not control one global family across all protocol epochs. These are local policy interventions, not additional whole-campaign or biological replications. Seeds and deployments are paired within initial states; no endpoint-driven top-ups or model exclusion occurs.

H.1 Naming intervention and format amendment

The naming experiment crosses visible gene names and opaque labels with TRUE and intact-tuple permutation. All conditions share randomized row order, a five-index response schema, 192 output tokens, three content attempts and outcome-blind shortage recovery. Only initially tested genes are eligible centres; their untested nearest neighbours form the acquired batch. The same score/hit tuple and row ordering are shown under named and opaque conditions. Full raw prompts, payloads, responses and execution records are retained. The first pilot used a parser that rejected quoted integer indices; V2 accepts digit-only integer tokens, including quoted/leading-zero integers, and reran all states before any acquisition endpoint was inspected. Decimals, signed indices, booleans and out-of-range values are rejected. The retained V1 cost is included. The format amendment changes the parser, not the score/hit rule or selected population.

Uniform-hit and first-hit naming pairs reproduce identical selections and executed batches in all 3,360 reference pairs. Every one of the 3,360 LLM decisions was reparsed against its transmitted prompt and its selected and recovered indices before outcome compilation. Reference sampling uses Python’s seeded sampling without replacement, with the frozen state-derived seed shared across naming conditions. This is not a permutation of physical truth labels: only the feedback shown to the controller changes; final scoring uses the unchanged released truth.

Table 22: Naming control: hits in one new native-size batch. B128 primary name and interaction intervals are 97.5%; other intervals are supporting 95%. B32 is never pooled into B128.

Contrast	B128 (720 states)	B32 (120 states)
Named minus opaque, TRUE	−0.93 [−1.33,−0.53]	−0.23 [−0.43,−0.05]
Naming-by-feedback interaction	−0.86 [−1.43,−0.28]	−0.35 [−0.57,−0.14]
TRUE minus permuted, named	2.81 [2.30,3.31]	1.07 [0.82,1.33]
TRUE minus permuted, opaque	3.67 [3.20,4.13]	1.42 [1.18,1.68]
Named LLM minus uniform-hit	−1.24 [−1.67,−0.82]	−0.45 [−0.71,−0.20]
Opaque LLM minus uniform-hit	−0.32 [−0.70,0.07]	−0.22 [−0.45,0.03]
Named LLM minus first-hit	−1.10 [−1.52,−0.69]	−0.44 [−0.69,−0.19]

Table 23: Naming-control heterogeneity and complete actor responses in B128. Each model has 120 matched states. Completion columns give the number accepting five actor indices out of 120; T/P denotes separate TRUE and permuted counts, not a ratio; recovery remains in all primary contrasts. Model intervals are supporting 95%.

Model	Named minus opaque, TRUE	Named T / P	Opaque T / P
Llama3.1 8B	−1.88 [−2.93,−0.91]	120/120	120/120
Qwen2.5 7B	−0.69 [−1.68,0.29]	118/116	120/120
Gemma3 4B	−1.29 [−2.29,−0.28]	118/117	120/120
Ministral8B	−0.73 [−1.60,0.14]	120/120	120/120
Qwen3 8B	−0.98 [−1.76,−0.24]	120/119	120/120
Phi-4 14B	0.03 [−0.18,0.24]	11/10	10/10

The selected population with complete actor responses in all four naming conditions contains 597/720 B128 states in 32 nonempty model–screen cells. Its named-minus-opaque contrast is −0.90

[-1.32,-0.50], and its naming-by-feedback interaction is -0.70 [-1.27,-0.12] (supporting 95%). Selection is post-treatment and changes the cell mixture; it cannot identify an unbiased actor-only effect. The deployment-inclusive primary result keeps every state. Mean action Jaccard overlap between names and opaque labels is 0.4308 under TRUE and 0.4269 under permutation. Naming changes tokenization and available semantic content together.

H.2 Implementation fidelity and independent executor check

The five-centre nearest-neighbour architecture is inherited, while our parser, legal-centre menu, stable tie order, integer quotas, deduplication and recovery are explicit reconstruction choices. The retained upstream artifacts do not supply an executable native implementation with every hyperparameter, so native score equivalence is not asserted. Native BDA search and critic tools are not part of this reconstruction. A separate Python sorting/quota/dedup implementation reproduces all 28,800 acquired slots in 252 decisions: every six-model/seven-screen cell, first and last fixed start IDs (5001 and 5020), initialization and the first LLM/TRUE and uniform/TRUE adaptive decisions. It uses the same released feature matrix and dot-product metric. This checks our algorithmic implementation, not a second representation family or an upstream reproduction.

H.3 Uncertainty in classical configuration selection

The retrospective exploratory check enumerates all $4^4 = 256$ ordered bootstrap samples of four development starts within each original screen. Candidate outcomes remain paired within a sampled start, and the original grid-order tie rule selects among all 20 settings. All 480 retained trajectories reproduce the frozen means and original winners. An independent multinomial-weight calculation reproduces the same selection frequencies. These 256 resamples are probability mass of an empirical bootstrap, not independent experiments or posterior probabilities of optimality. No final-test setting, outcome or interval is changed. The transferred CAR-T/external setting has no screen-specific development sample. Four development starts limit this uncertainty estimate.

Table 24: Exploratory development-selection uncertainty. A setting with positive bootstrap frequency need not be a plausible winner in a new biological source. Candidate IDs refer to the retained fixed GP/linear/observed-score grid in Appendix I.

Screen	Original setting	Reselection (%)	Settings ever selected
Carnevale	C00	34.8	5
Schmidt IFNG	C02	93.8	3
Schmidt IL2	C00	55.9	6
Sanchez	C13	9.4	6
Sanchez-down	C01	71.1	3
Scharenberg	C00	54.3	3

The full candidate-by-screen frequency table and paired development hit counts accompany the numerical evidence. The primary observed-hit controller contrasts do not use this classical selection.

H.4 Output-budget and response-format factorial

The open-centre test keeps the original legal pool and system/user prompts, then crosses 384 versus 1,536 output tokens, ORIGINAL versus SOLUTION_ONLY, and TRUE versus WITHIN_BATCH_PERMUTED. SOLUTION_ONLY appends an instruction to omit Reflection and Research Plan and return exactly five legal HGNC centres, repeating the existing neighbour-execution contract. Temperature is 0.2, Qwen3 thinking is disabled, and model/provider routes are unchanged. Seeds match budget/format comparisons within each state, feedback arm and content-attempt index. The original parser accumulates legal unique centres across up to three content attempts; outcome-blind recovery fills any shortage. The eight condition order is frozen per state. The non-LLM TRUE/permutated references reuse the original first adaptive decision, verified against identical tested genes, feedback packets, acquired batches and released outcomes. Reusing a reference across four interface conditions does not create additional stochastic replicates.

1154 Transport resumptions retained the frozen population, payloads, seeds, parser and outcomes. The
1155 physical retry bound increased from 12 to 36 per content attempt; rejected requests and their con-
1156 servative charges remain retained. Collection pauses did not change scientific conditions. No acqui-
1157 sition endpoint was inspected until all 840 states and 6,720 LLM decisions were complete. Detailed
1158 dated transport and storage records remain in the evidence archive.

Table 25: Open-centre output-budget factorial: four B128 primary 98.75% intervals; all B32 and additional contrasts supporting 95%. Values are hits in one new batch, not full-campaign gains.

Contrast	B128 (720 states)	B32 (120 states)
1,536 minus 384, original TRUE	0.11 [-0.39,0.64]	0.04 [-0.29,0.38]
Solution-only minus original,384 TRUE	0.50 [-0.14,1.14]	0.20 [-0.18,0.57]
LLM minus uniform,1,536 solution TRUE	-1.63 [-2.24,-1.05]	-1.07 [-1.42,-0.72]
Differential feedback utility,1,536 solution	-1.40 [-2.19,-0.64]	-1.39 [-1.78,-0.99]
Budget-by-format interaction, TRUE	0.03 [-0.46,0.50]	0.09 [-0.32,0.50]
1,536 minus 384, solution TRUE	0.14 [-0.15,0.44]	0.13 [-0.12,0.39]
Solution-only minus original,1,536 TRUE	0.53 [0.02,1.03]	0.29 [-0.05,0.65]
LLM minus uniform,384 original TRUE	-2.27 [-2.78,-1.77]	-1.40 [-1.76,-1.04]
LLM minus first,1,536 solution TRUE	-0.92 [-1.39,-0.47]	-0.99 [-1.34,-0.64]

Table 26: All eight B128 interface conditions: mean next-batch hits and supporting 95% interval; complete actor decisions out of 720; physical requests; recorded/conservative API cost in USD.

Condition	Hits	Complete	Requests	Cost
384 Orig T	12.60 [12.13,13.07]	660	1125	0.2214
384 Orig P	10.68 [10.24,11.13]	676	1100	0.2160
384 Sol T	13.10 [12.58,13.62]	708	984	0.1805
384 Sol P	10.29 [9.87,10.73]	693	996	0.1729
1536 Orig T	12.71 [12.26,13.17]	704	1017	0.2078
1536 Orig P	10.78 [10.36,11.22]	711	1048	0.2236
1536 Sol T	13.24 [12.73,13.75]	706	974	0.1752
1536 Sol P	10.33 [9.89,10.80]	695	1045	0.1995

Table 27: Model-specific B128 controller and feedback contrasts at 1,536 tokens/solution-only; supporting 95% intervals,120 states/model. All eight actor-completion counts follow in the text.

Model	LLM minus uniform, TRUE	Differential feedback utility
Llama3.1 8B	0.77 [-0.29,1.84]	-0.52 [-1.92,0.86]
Qwen2.5 7B	0.52 [-0.48,1.52]	0.46 [-0.90,1.79]
Gemma3 4B	-1.38 [-2.37,-0.42]	-0.83 [-2.06,0.38]
Ministral8B	-5.98 [-7.05,-4.93]	-4.87 [-6.38,-3.41]
Qwen3 8B	-0.87 [-2.37,0.58]	-0.38 [-2.10,1.29]
Phi-4 14B	-2.81 [-3.96,-1.68]	-2.27 [-4.08,-0.48]

Table 28: Complete actor decisions out of 120 per model and condition. O: original response; S: solution-only; T: TRUE feedback; P: within-batch permutation. Recovery fills shortages.

Model	384 tokens				1,536 tokens			
	O/T	O/P	S/T	S/P	O/T	O/P	S/T	S/P
Llama3.1 8B	116	118	120	116	118	119	120	118
Qwen2.5 7B	119	117	113	105	116	120	110	106
Gemma3 4B	117	120	118	115	119	119	119	114
Ministral8B	78	92	120	120	120	119	120	120
Qwen3 8B	119	119	120	120	120	119	120	120
Phi-4 14B	111	110	117	117	111	115	117	117

The all-eight-complete sensitivity retains 569/720 states in 36 nonempty cells. Its LLM-minus-uniform contrast is -1.63 $[-2.17, -1.09]$ and differential feedback utility is -1.42 $[-2.13, -0.72]$ (supporting 95%). This is a selected population with changed cell weights, not an unbiased actor-only effect.

The open-centre programme retains 10,700 physical request records at a conservative total of \$1.8226, including rejected-request charges and one interrupted reservation without a saved response or settlement. That reservation remains fully charged but is not credited as a verified physical request or observation. All 6,720 LLM decisions were reparsed, all 3,360 reference decisions reproduced, and every acquisition batch validated for legality, uniqueness, prior-tested exclusion and released-truth scoring. The one-step factorial does not establish the performance of 1,536-token solution-only policies over evolving full campaigns, nor under a separately optimized random-executor prompt.

H.5 Interrupted truthful-instruction check

A further check was specified to separate executor instructions from permission to propose new centres. It used Gemini3.8Flash with native JSON, low reasoning, temperature 1 and a 4,096-token output limit on 20 existing LLM-initialized starts: ten each on Carnevale and IL2. These were exposed conditions, not new biological validation. Four LLM continuations crossed NN execution with unrestricted or observed-only centres, and random execution with NN-oriented or accurate random-executor instructions. Three uniform-hit rivals matched executor and centre eligibility. Each campaign targeted 640 tests, including its 128 shared initial tests. All adaptive histories were generated afresh. Observed-only allowed any previously acquired gene as a centre; it did not restrict the LLM to known hits. Uniform-hit used up to five observed hits, completing any remaining slots by outcome-blind hash order within the allowed centre pool. Each numerical condition required one such completion, on the same low-hit initialization. The two NN numerical conditions consequently have identical centres and batches in 76/80 decisions, not all 80. The opportunity contrast subtracts their actual yields; it does not assume they are identical.

Random execution retained untested centres that NN expansion would include, then sampled remaining slots uniformly. The accurate prompt described this rule. If all centres must already have been observed, none can be newly acquired: with matched random seeds and history, random filling then cannot depend on the chosen centres. This is a property of the executor, not an additional measured LLM success or failure.

Provider failures left 18 of 20 starts complete across all seven conditions (eight Carnevale, ten IL2), despite a resumption retaining the same model, prompts and planned starts. Terminal campaign counts were 19 unrestricted-NN LLM, 20 fixed-instruction random LLM, 18 accurate-instruction random LLM, 20 observed-only NN LLM, and 20 for each uniform-hit condition. The 550 completed adaptive decisions include 310 LLM decisions; one required centre recovery. Of 335 physical requests, 320 returned successful responses and 15 were transport/error records. No missing response was replaced by another model. All earlier successes, failures, charges and additional transport retries are retained in the collection record; the source archive states the resumption contract. The last recovery attempt stopped on its first further provider error; no successful decision or planned start was discarded.

The planned primary comparisons were the instruction effect under random execution, the unrestricted controller gap under NN, and the change in that gap when centres must be observed. We do not treat completed starts as an unbiased replacement for the planned panel or report confirmatory confidence intervals from them. Instead, every unfinished trajectory retains its observed hit count; its remaining distinct tests can add between the logically required minimum and the smaller of the remaining test budget and remaining eligible hits. These finite-pool bounds give the following ranges for the planned 20-start mean differences:

	Contrast (hits/640)	Mean bounds	Bootstrap envelope
	Accurate minus NN-oriented instructions, random executor	[1.15, 39.55]	[−4.40, 95.12]
	LLM minus uniform-hit, unrestricted NN	[−14.85, 10.75]	[−35.00, 77.12]
	Observed-only minus unrestricted NN controller gap	[−19.90, 5.70]	[−86.79, 27.80]

Mean bounds describe the fixed planned panel, not a population confidence interval. The instruction bound stays positive even at the least favourable completion, but this does not include variation across starts. To assess that sensitivity retrospectively, we apply the planned 20,000 within-source paired bootstrap draws to each start’s lower and upper bounds. For each resample, we compute means using all lower bounds and all upper bounds. Taking the lower endpoint from the first distribution and the upper endpoint from the second encloses every compatible completed-data 98.333% percentile interval. This follows because resampled means and their quantiles are monotone in each outcome. The envelopes are outer sensitivity bounds, not calibrated coverage guarantees; they inherit the bootstrap approximation and condition on these two selected sources. All three permit both signs. Thus the positive fixed-panel instruction bound is not a confirmed instruction benefit once missingness and start variation are considered together.

The source breakdown also matters. Carnevale’s instruction mean can range from −6.00 to 70.80; IL2’s completed mean is +8.30. Their unrestricted NN controller differences are [−2.20, 49.00] and −27.50, respectively; the opportunity interactions are [−43.90, 7.30] and +4.10. Point values from the complete source are descriptive, not a substitute for finishing the planned panel or an independent biological replication.

These records preserve an attempted falsification without adding a completed confirmation to the main evidence. The source archive includes the per-start bounds, source breakdown, bootstrap-envelope replay and outcome-status table. Full collection records remain preserved separately. The explanation of executor dependence remains unresolved.

I Evolving executor campaigns and development-selected references

Prospective question and scope. A frozen-history change of executor measures the immediate value of expansion at a state reached by NN. It cannot predict the history reached after four random-expansion rounds. The fresh experiment therefore crosses LLM/uniform-hit with NN/random expansion on new starts 4001–4020 for each of Qwen2.5 and Ministral on all six screens. These identifiers and generation namespaces are disjoint from the original 600-start panel. Every block contains one shared initial NN batch, then four adaptive trajectories. They use TRUE feedback and their own acquired histories. The protocol and runner were hashed before the first calls. A technical pilot of two blocks was retained as part of the fixed 240-block panel after confirming legal execution and successful transport, without using biological yield to admit models or select sample size. The four predeclared outcome patterns were positive effects for both controllers, a controller-specific effect, no clear effect, or reversal. Mixed signs remain in the results. No significance-driven extra starts were allowed.

Exact random expansion. At each arm’s current history, the NN compiler first determines which of the five operational centres would themselves occupy acquired slots. Random expansion includes exactly this centre subset. It fills the remaining batch uniformly without replacement from untested genes excluding all proposed centres. It preserves budget and centre inclusion at that local intervention while allowing future candidate pools, observations and proposals to diverge. When uniform-hit selects previously observed centres, none can enter the new batch. An LLM can instead propose a

1246 novel centre that is retained. Thus random expansion is not a pure random policy, and LLM versus
1247 uniform-hit under that executor includes the consequences of their different centre choices. The
1248 identical NN-oriented prompt is used under both executors. This fixed-policy intervention does not
1249 compare two separately prompt-optimized controllers or isolate an internal reasoning mechanism.

1250 **Common sampling and operational outcome.** Controller-executor order is independently shuf-
1251 fled within a block. Generation seeds are paired across executor arms by model, screen, start, round
1252 and attempt; providers do not guarantee exact seed reproducibility. Temperature 0.2, the 384-token
1253 output limit, three-attempt parser, outcome-blind recovery, legal universe and hosted routes are in-
1254 herited unchanged. Missing actor proposals stay in the operational policy. There are 240 initial
1255 and 1920 adaptive logical LLM decisions, rather than four independent copies of each initial de-
1256 cision. Each of the 960 adaptive-policy trajectories has five rounds; the 240 classical confirmation
1257 trajectories share the same initial batches. B640 and B160 yield totals are never pooled in raw units.

1258 **Disjoint development selection.** Four random-initialization development starts 3001–3004 per
1259 screen are shared by twenty fixed candidate settings. Twelve binary-hit GP settings cross length
1260 0.25/0.5/1, noise 0.01/0.1 and exploration multiplier 1/2. Six binary-hit linear settings cross noise
1261 0.1/1/10 and multiplier 1/2. Two settings use the original observed-only absolute standardized score
1262 target: GP length 0.5, noise 0.01, multiplier 2 and linear noise 1, multiplier 1. All fit only previously
1263 observed outcomes; whole-screen moments are confined to evaluation. GP covariance is the same
1264 cosine-RBF kernel as the original experiment with the specified length. Linear covariance is the
1265 dot product on unit-normalized features. Within a batch, unobserved outcomes are provisionally
1266 replaced by their posterior-mean predictions (“fantasies”), with exact incremental covariance updates.
1267 A direct full-posterior refit independently verified the selected batch for a nonlegacy GP setting;
1268 legacy settings reproduced the previous implementation on a synthetic case.

1269 One rule per screen maximizes average final development hits, with the first grid member breaking
1270 a tie. All 480 development trajectories and losses are retained. Selection was frozen at 20:00:16
1271 UTC on 7 September 2026, after development completed and before confirmation outcomes were
1272 inspected. Confirmation acquisition was already running under the frozen protocol; its files were
1273 not inputs to the selection function. The selected configurations are applied unchanged to 240
1274 fresh shared LLM initial batches. This is same-screen stochastic confirmation with an initialization-
1275 distribution shift, not a held-out biological-source claim. The twenty settings do not count as twenty
1276 independent baseline families. Additional development computation is explicit and is not matched
1277 to the LLM’s inference cost.

1278 **Analysis and verification.** For each controller, paired NN-minus-random cumulative-hit differ-
1279 ences are averaged with equal weight across ten model–screen cells at B640. Twenty thousand
1280 stratified start bootstraps provide two 97.5% primary percentile intervals. The B160 panel, per-
1281 cell estimates, controller contrasts, interaction and alternate endpoints use supporting 95% intervals
1282 without family-wide significance claims. The earlier matched-controller study had its own two-
1283 test primary family; the subsequent six-model balanced panel has the four-test family defined in
1284 Appendix M. These studies retain their distinct populations and families. The reported intervals
1285 condition on these fixed screens, configurations and selected classical settings; they do not integrate
1286 uncertainty from repeating development selection. No equivalence margin or universal incapability
1287 inference is introduced. Full raw replay checks every batch against released outcomes, recomputes
1288 both executors, verifies TRUE feedback tuples and common initializations, and reconstructs the cu-
1289 mulative endpoints. The distributed no-credential replay rebuilds all 1200 fresh trajectories and the
1290 two primary intervals from retained ledgers; feature-neighbour replay additionally needs the hash-
1291 identified upstream feature matrix.

Table 29: Development-selected configurations. Selection uses four random-initialization starts per screen and mean final released hits. Every selected setting fits binary observed-hit labels. Development means are selection data, not confirmation effects.

Screen	Kernel	Length	Noise	β	Dev. hits
Carnevale	GP	0.25	0.01	1.0	33.75
Sanchez	LINEAR	–	0.1	2.0	38.25
IL2	GP	0.25	0.01	1.0	35.25
IFNG	GP	0.25	0.1	1.0	51.50
Sanchez-down	GP	0.25	0.01	2.0	55.25
Scharenberg	GP	0.25	0.01	1.0	18.25

Table 30: Per-cell executor effects with supporting 95% paired-start intervals. Each cell contains 20 starts. Scharenberg uses 160 tests; all other screens use 640.

Model/screen	Ω_{LLM}	Ω_{uniform}
Qwen2.5/Carnevale	+26.65 [21.65, 31.05]	+25.95 [18.95, 31.80]
Qwen2.5/Sanchez	+17.50 [13.05, 21.60]	+16.85 [13.35, 20.40]
Qwen2.5/IL2	+103.90 [85.55, 118.70]	+135.05 [124.90, 142.85]
Qwen2.5/IFNG	+47.15 [35.55, 58.50]	+67.35 [62.65, 72.00]
Qwen2.5/Sanchez-down	+7.30 [4.75, 10.05]	+3.15 [-0.35, 6.55]
Qwen2.5/Scharenberg	+4.20 [2.50, 5.80]	+7.30 [6.20, 8.55]
Ministral/Carnevale	+19.90 [16.05, 23.30]	+26.90 [23.15, 30.10]
Ministral/Sanchez	+13.55 [9.85, 17.20]	+20.10 [15.25, 24.75]
Ministral/IL2	+84.60 [70.60, 97.80]	+140.70 [137.90, 143.50]
Ministral/IFNG	+34.65 [29.95, 39.30]	+66.35 [62.15, 70.30]
Ministral/Sanchez-down	+6.60 [4.00, 9.05]	+10.95 [8.10, 13.80]
Ministral/Scharenberg	+1.45 [0.10, 2.80]	+5.90 [4.70, 7.15]

Table 31: Fresh generation ledger. Complete/partial/zero are logical decisions by accepted actor-centre count. Attempts include every retained retry and transport result. Raw responses, prompt bytes and request identifiers remain available.

Model	Decisions	Attempts	Complete	Partial	Zero	Max input
Ministral	1080	2600	582	230	268	8034
Qwen2.5	1080	1622	989	84	7	7982

Table 32: Alternate-endpoint executor effects. Whole-screen standardization is used for evaluation only. These supporting 95% intervals do not replace the primary released-hit family.

Panel/endpoint	Ω_{LLM}	Ω_{uniform}
B640/Strict hits	+36.18 [33.54, 38.66]	+51.43 [49.87, 52.89]
B640/Mean absolute z	+0.18 [0.17, 0.19]	+0.25 [0.25, 0.26]
B160/Strict hits	+1.35 [0.43, 2.25]	+4.40 [3.60, 5.23]
B160/Mean absolute z	+0.05 [0.03, 0.08]	+0.13 [0.10, 0.15]

Development-grid availability. The full 20-setting, six-screen development matrix is retained in the full archival manuscript and its source archive. Table 29 reports every selected setting; the following sensitivity analysis reports whether selection changes when a development start is omitted. The selected settings and confirmation results are unchanged.

Resource accounting. The fresh experiment incurred \$1.602610 in charged or conservatively reserved hosted requests, including failures. No new GPU job was used. Local development and

1298 classical confirmation used CPU computation, reported as elapsed runtime. These costs cover this
1299 experiment rather than the entire project.

1300 **Supporting model-specific contrasts.** These exploratory decompositions were computed after
1301 the pooled results and review, from all 240 completed blocks. They neither change the primary
1302 family nor add model replications. Within each model and budget, 20,000 paired start-bootstrap
1303 replicates resample within screen; B640 averages five screens equally and B160 contains Scharen-
1304 berg only. Intervals condition on the fixed screens and model routes. Table 33 reports hits per stated
1305 budget; the interaction is the NN controller contrast minus the random-expansion contrast.

Table 33: Model-specific supporting 95% intervals. Each B640 row contains 100 starts and each B160 row 20. All contrasts are paired within starts.

Model	Budget	LLM—uniform, NN	LLM—uniform, random	Interaction
Qwen2.5	640	-9.31 [-13.76, -5.25]	-0.14 [-1.39, 1.12]	-9.17 [-13.86, -4.92]
Qwen2.5	160	-2.00 [-2.90, -1.15]	1.10 [-0.55, 2.95]	-3.10 [-5.15, -1.15]
Ministral	640	-20.64 [-23.78, -17.57]	0.50 [-0.88, 1.89]	-21.14 [-24.55, -17.70]
Ministral	160	-4.75 [-6.05, -3.50]	-0.30 [-1.25, 0.80]	-4.45 [-6.00, -2.90]

1306 **Development-selection sensitivity.** For each screen we repeat the original grid ranking four times,
1307 omitting one of the four random-initialization development starts. The objective and first-grid-
1308 member tie-break are unchanged; confirmation outcomes are never used. The table records how
1309 often the original setting remains the winner and its full-development mean-hit lead over the runner-
1310 up. This small-sample sensitivity check describes instability, not the probability that a setting is
1311 optimal. It does not replace the frozen selected comparator or adjust its conditional intervals.

Table 34: Supporting leave-one-development-start-out sensitivity. Winner identifiers follow the original twenty-setting grid. Gap units are final hits at the screen’s original budget.

Screen	Selected	Retained/4	Omission winners, seed order	Runner-up gap
Carnevale	C00	2	C00, C13, C14, C00	0.50
Sanchez	C13	0	C07, C00, C19, C19	0.25
Sanchez↓	C01	4	C01, C01, C01, C01	1.25
IL2	C00	3	C00, C00, C02, C00	0.75
IFNG	C02	4	C02, C02, C02, C02	1.50
Scharenberg	C00	2	C02, C00, C00, C18	0.25

Table 35: Eligible acquisition populations. Coverage is the fraction of scored genes retained after the feature intersection; hit prevalence uses the eligible pool and released hit arrays. Directions from one study remain dependent.

Screen	Scored	Eligible	Coverage (%)	Hits (%)
Carnevale	18,861	17,136	90.9	858 (5.01)
Sanchez	18,469	17,204	93.2	859 (4.99)
Sanchez↓	18,469	17,204	93.2	883 (5.13)
IL2	18,939	17,758	93.8	634 (3.57)
IFNG	18,418	17,009	92.3	861 (5.06)
Scharenberg	1,055	1,013	96.0	48 (4.74)

1312 J Earlier tests and unresolved predictions

1313 These earlier cohorts do not add observations to the current balanced panel. In the earlier matched
1314 campaign, LLM-minus-uniform yield was +0.31 [−2.05, 2.69] hits/640 and the feedback interac-
1315 tion was −2.53 [−5.85, 0.84] (97.5% intervals). A separate first-adaptive comparison gave +1.025

1316 $[-0.041, 2.046]$ hits/128; its composite included post-outcome provider replacement and is not un-
 1317 touched confirmation. The distinct starts and horizons must not be subtracted as a paired effect.
 1318 A matched 80-state layout check gave a feedback-by-layout interaction of $+1.12 [-0.61, 2.86]$
 1319 hits/128 (95%); it did not establish that either display format is universally neutral. Full cohort
 1320 contracts, tables and plots remain in the preserved full version.

1321 **J.1 Earlier horizon prediction and measurement limitations**

1322 The earlier E1 study used 45 model–screen–run blocks, with early and late fixed states. E2 used 40
 1323 paired four-arm blocks on two screens and two configurations, with ten runs per cell. The govern-
 1324 ing E2 prediction was that the WITHIN-minus-NO contrast would grow between rounds two and
 1325 five. Its interaction was $+0.275 [-2.400, +3.325]$ hits per128 tests; the cumulative difference was
 1326 $+1.600 [-5.125, +8.725]$ hits per640 tests. Neither the prediction nor its equivalence criterion was
 1327 met. Equivalence required the full 95% interval within ± 1 hit/B128 or ± 5 hits/B640, respectively.

1328 The secondary TRUE-minus-WITHIN difference was positive (retained interval $[0.550, 9.900]$,
 1329 Holm $p = 0.01263$), while stricter and continuous endpoints did not show the same clear sepa-
 1330 ration. This does not rescue the failed governing prediction. E1 included 357 all-out-of-memory
 1331 decisions, concentrated in late outcome-bearing arms; late NO decisions generated successfully. E2
 1332 had 31 all-out-of-memory decisions, all in Qwen2’s final round. Recovery supplied actions when
 1333 generation failed. These comparisons therefore concern the executed pipelines, not a clean model-
 1334 memory effect. Excluding recovery cases would select on a post-treatment event.

1335 The full archival manuscript retains the cell tables, all endpoint sensitivities, original protocols and
 1336 selected raw-response examples. They are historical evidence, not additional replications of the
 1337 current balanced experiment.

1338 **K Sources and reproducibility**

1339 **K.1 Screens, outcomes and feature provenance**

1340 IFNG and IL2 derive from the cytokine-screen resource of Schmidt et al. [2022]; Carnevale concerns
 1341 T-cell proliferation under adenosine agonism [Carnevale et al., 2022]; Sanchez and Sanchez-down
 1342 concern neuronal tau expression [Sanchez et al., 2021]. We use the task strings, processed out-
 1343 come CSVs and hit arrays retained in the BDA repository, rather than claiming to reprocess the
 1344 original experimental count matrices. These are five tasks from three study lineages. Sanchez and
 1345 Sanchez-down share score ancestry and pool support; they are different hit authorities, not inde-
 1346 pendent biological replications. Sanchez is held out from E1 as a task variant, not an independent
 1347 source from Sanchez-down.

1348 **Primary and sensitivity outcomes.** The first ground-truth CSV column is mapped to uppercase
 1349 gene symbols and the second to floating-point scores. Primary hits are uppercase membership in
 1350 `topmovers_SCREEN.npy`; each selected gene is scored against that exact array. The released IL2
 1351 hit count differs from a uniform top-five-percent construction, so primary membership must not be
 1352 reconstructed from a generic quantile rule. Continuous outcomes use $|z_g| = |s_g - \mu_s|/\sigma_s$, with mean
 1353 and population standard deviation computed over the complete processed screen CSV. Sensitivity
 1354 counts use $|z_g| \geq 1.5, 2.0, 2.5$. The manuscript does not replace the primary hit authority with these
 1355 thresholds.

1356 **Feature and pool construction.** The retained Achilles resource is the file fetched by BDA’s
 1357 released downloader at Figshare file ID 49843176 [BioDiscoveryAgent contributors, n.d.,
 1358 Broad Institute, n.d.]. It has 18,119 gene columns and 808 cell-line rows, with SHA-256
 1359 `71f32e3515dfa24d3bff559ea99fd5a0fa102c264064a6377e13d8eecd54298f`. Gene names
 1360 are taken from the prefix before the parenthesized identifier. Rows with any missing feature value
 1361 are dropped, gene vectors are transposed to float32 and normalized to unit Euclidean norm. Candi-
 1362 date pools intersect feature genes with screen-scored genes in sorted order. The E1/E2 feature cache
 1363 and each pool have separate manifest hashes. A citation to the current DepMap portal identifies
 1364 the resource family, not a verified release number for this retained CSV. Its redistribution authority
 1365 remains unresolved, so the review snapshot omits the feature matrix and identifies the exact retrieval
 1366 and hash instead.

1367 K.2 Current-panel prompt examples

1368 These are the recorded system and initial user messages for Qwen2.5, Carnevale, start5001, followed
 1369 by its first-adaptive user message. Original wording, including repeated phrases, is preserved: this
 1370 is the tested interface, not a proposed improved prompt. The two gene-score tables in the adap-
 1371 tive message are omitted only in this display; the complete transmitted messages and hashes are
 1372 in `synthesis_reproduction/CURRENT_PROMPT_EXAMPLES.json`. The example used tempera-
 1373 ture 0.2, 384 output tokens and the Phala route. Later response-contract variants are specified in
 1374 Appendix A.3; the system/task definitions vary by source as recorded in Table 5. Historical E1/E2
 1375 templates remain in the full version and the retained source files.

1376 You are a biomedicine expert who will assist me on problems in drug discovery. I am
 1377 planning to run a CRISPR screen to identify genes that identify genes that,
 1378 upon being knocked out, would boost the efficacy of engineered T cells in the
 1379 presence of an adenosine agonist that creates an immunosuppressive condition. I
 1380 can only perturb exactly 128 genes at a time. For each predicted perturbation,
 1381 I am able to measure out the the change in T cell proliferation which will be
 1382 referred to as the score. I can only do 5 rounds of experimentation. After
 1383 every round of experiment, I will provide you with feedback on your predictions
 1384 , including the correctly identified genes called hits and the corresponding
 1385 score. The predictions which are not hits will be included in other results.

1386 This is round 1. We are beginning with our experiments.
 1387 Here is a strategy to follow: Choose genes that are very different in their
 1388 biological pathways to discover what pathways give you hits.
 1389 Please propose 5 different yet valid gene names as per the HGNC nomenclature you
 1390 want to explore next. Note that I will choose unexplored genes closest to your
 1391 predicted genes to form the predictions. Your response should exactly follow
 1392 the format:
 1393 ****Reflection:** Thoughts on previous results and next steps.
 1394 ****Research Plan:** The full high-level research plan, with current status and
 1395 reasoning behind each proposed approach. It should be at most 5 sentences.
 1396 ****Solution:**
 1397 **##** <Gene 1>
 1398 **##** <Gene 2>
 1399 ...
 1400 **##** <Gene 5>
 1401 Each gene in the solution should only be the gene name in the HGNC nomenclature. DO
 1402 NOT ADD ANY COMMENTS IN THE SOLUTION OR AFTER THE SOLUTION.

1403 This is round 2.
 1404 Here is the feedback on all your predictions till now:
 1405 [HITS]
 1406 [Recorded hit gene-score rows omitted in this display]
 1407 [OTHER RESULTS]
 1408 [Recorded remaining gene-score rows omitted in this display]
 1409 Here is a strategy to follow: Update your priors appropriately and choose genes that
 1410 gave you hits. Also, be sure to explore by including some genes that could
 1411 give hits.
 1412 Please propose 5 different yet valid gene names as per the HGNC nomenclature you
 1413 want to explore next. Note that I will choose unexplored genes closest to your
 1414 predicted genes to form the predictions. Your response should exactly follow
 1415 the format:
 1416 ****Reflection:** Thoughts on previous results and next steps.
 1417 ****Research Plan:** The full high-level research plan, with current status and
 1418 reasoning behind each proposed approach. It should be at most 5 sentences.
 1419 ****Solution:**
 1420 **##** <Gene 1>
 1421 **##** <Gene 2>
 1422 ...
 1423 **##** <Gene 5>

Each gene in the solution should only be the gene name in the HGNC nomenclature. DO NOT ADD ANY COMMENTS IN THE SOLUTION OR AFTER THE SOLUTION.

K.3 Reproduction and archive scope

The anonymous manuscript-source package contains all included TeX tables, literal prompt renderings and response-summary tables, bibliography, figures and the official style. The separate retained evidence snapshot contains original E1/E2 raw requests/responses and ledgers, protocols, analysis files, decision/attempt linkage, control checks, sensitivity exports, matched traces, the first-adaptive matched-reference records, and the complete new matched-panel records and reconstruction instructions. Historical records can contain superseded interpretations; the current interpretation and analysis version are identified separately. Retention in a local handoff archive is not a claim of publicly accessible release.

The feature matrix is omitted because upstream redistribution authority is unresolved. The exact source URL and hashes permit its identification. Source-scale raw execution has a different archive scope from the complete E1/E2 evidence; included source summary and provenance tables should not be described as a complete raw source-model release. The paper-source package builds with `latexmk -pdf -interaction=nonstopmode -halt-on-error main.tex`. Evidence reconstruction commands and dependency scope are documented in that archive; the retained-data scripts run without loading any model weights.

Reproducibility limits. An independent E1/E2 raw-to-ledger audit reconstructs parser attempts, recovery centres, exact gene/score/hit authority and E2 endogenous history. Recomputing standardized scores differs by at most 4.44×10^{-16} across observed environments, under a recorded 10^{-12} tolerance; primary scores and hit membership match exactly. Earlier cross-BLAS compiler replay found 24 near-tie order swaps but identical sets and outcomes in all 800 E2 ledgers. Set/outcome reconstruction is therefore distinguished from bitwise rank identity and from identifying a causal effect of the compiler itself.

K.4 Scope, rights and reporting

This work analyzes public processed screens and generates no new laboratory interventions, patient decisions or biological efficacy evidence. Prospective use would need independent biological validation and oversight appropriate to the proposed experiments. Model outputs quoted as records are not endorsed biological recommendations.

The BDA repository’s license and the separate feature-resource redistribution boundary are recorded in the evidence archive. No model weights are redistributed; model licenses and provider conditions are identified separately. No author identity, credentials or identifying local path is intended for the anonymous submission surface.

L Relationship transport and predictive-boundary stress tests

These additions distinguish retrospective analysis of previously exposed outcomes from newly frozen intervention campaigns. They do not change earlier primary families or model/source totals. The analyses below test source-specific coexistence of feedback benefit and controller loss, prediction of the better complete policy, and feedback use under a separately collected external-screen contract.

L.1 Joint fixed-source evidence

The 720 B640 starts are resampled within model/readout cells, jointly for the three contrasts. Sources receive equal weight in the aggregate and cells equal weight within source. Twelve source-by-contrast intervals use Bonferroni coverage $1 - 0.05/12 = 99.583\%$. The aggregate uses

1470 $1 - 0.05/3 = 98.333\%$. These retrospective checks do not turn computational starts into biological
1471 replication.

Table 36: Fixed-source paired effects in hits/640. Source rows use 99.583% intervals; aggregate 98.333%.

Source	Controller	Assignment	Availability
Schmidt	−27.07 [−31.63,−22.52]	16.72 [10.30,23.11]	45.34 [39.59,51.00]
Carnevale	−2.65 [−5.06,−0.18]	5.17 [2.23,8.18]	14.32 [11.54,17.05]
Sanchez	−1.41 [−3.01,0.24]	1.42 [−0.14,3.04]	2.52 [1.04,3.99]
CAR-T	−0.24 [−0.75,0.29]	−0.20 [−0.67,0.25]	−0.96 [−1.56,−0.36]
Equal-source	−7.84 [−8.98,−6.71]	5.78 [4.27,7.26]	15.31 [13.96,16.65]

1472 L.2 Source-and-family-excluded prediction

1473 Six LLM-minus-uniform first-state features measure centre reuse, actor completeness, mean non-
1474 centre cosine distance, log neighbour rank, centre-pair cosine and observed-history nearest-eight-
1475 neighbour hit support. No unacquired outcome enters a feature. Features require both first proposals,
1476 so their acquisition is not free. The target is final whole-campaign LLM-minus-uniform yield/640.
1477 For each source/family pair, training excludes both groups; the two Qwen checkpoints share a fam-
1478 ily. All 720 B640 starts are predicted exactly once across 20 folds. Training-only weighted center-
1479 ing/scaling and ridge penalty 10 are fixed, without tuning. Training sources and their cells receive
1480 equal weight. Rivals use a training intercept, a geometry-only ridge and a fixed lower-distance rule.
1481 The hindsight oracle is unattainable.

1482 A test fold contains every start from one source group and one model family. The four source
1483 groups and five families form twenty folds, with 20–80 test starts each; both Qwen checkpoints
1484 stay together, as do paired readouts from a source. Training uses only the remaining original B640
1485 starts that share neither the held-out source nor its model family (320–500 starts). For example, a
1486 Schmidt–Qwen test fold excludes all Schmidt starts and all Qwen starts from training, not only their
1487 intersection. Preprocessing and equal-source fitting use training only. The rivals predict a constant
1488 fitted on training data or use mean neighbour distance alone. Intervals condition on fitted folds.
1489 These are held-out analyses of already studied sources, not confirmation on a previously untouched
1490 source.

Table 37: Retrospective full-campaign prediction. MSE improvement is intercept error minus six-
feature error (hits²); selection yield is the predicted winner minus always uniform (hits/640). Inter-
vals below the table condition on fitted folds, not training or source-population uncertainty.

Source	MSE improvement	Ridge selection	Geometry selection	Oracle
Schmidt	108.31	−2.60	−1.45	1.28
Carnevale	−69.07	0.33	0.00	2.63
Sanchez	−86.71	0.53	0.20	3.17
CAR-T	34.68	0.02	0.12	0.72
Equal-source	−3.20	−0.43	−0.28	1.95

1491 Equal-source MSE improvement is −3.20 [−19.84,13.42]; ridge selection gives −0.43
1492 [−0.82, −0.05], and full-minus-geometry selection −0.14 [−0.58, 0.29] (supporting 95%). A fixed
1493 199-permutation control shuffles outcomes within model/readout cells and refits the entire pipeline,
1494 preserving cell means. Its one-sided empirical $p = 0.005$ detects association beyond that null, but
1495 the observed fit still does not improve overall prediction risk. Static choice between saved complete
1496 policies is not online routing. These features therefore do not establish a reliable rule for choosing
1497 the higher-yield controller.

1498 L.3 Fresh external feedback interventions

1499 The source and original external controller outcomes were already exposed. A new pre-call contract
1500 freezes Qwen2.5-7B/Phala and Phi-4/DeepInfra bf16, 20 starts each (91001–91020), temperature

0.2, 384 output tokens, three logical attempts, full history and outcome-blind recovery. All six controller/feedback arms share one LLM initial 128-test batch per start, then follow their own four adaptive 128-test rounds under NN execution. No old TRUE trajectory is combined with a new feedback arm. All 240 trajectories completed; 1,000 unique decisions were replayed against source truth and executor, with 960 feedback packets checked. The decision count includes one shared initialization and six four-round continuations for each start: $40(1 + 6 \times 4) = 1,000$.

Table 38: Supporting model-specific external effects (hits/640; 95% intervals), separate from the pooled four-contrast 98.75% family reported in Table 16.

Contrast	Qwen2.5-7B	Phi-4
Controller	-11.90 [-25.75,1.80]	-55.20 [-86.90,-26.60]
Assignment	19.35 [0.10,38.85]	2.60 [-33.90,35.65]
Availability	28.65 [8.35,47.40]	89.05 [53.35,121.65]
Feedback interaction	-64.45 [-96.35,-32.65]	-79.20 [-127.65,-33.65]

Actor completion is 507/520 logical LLM decisions; recovery is included throughout. The 745 requests include four synthetic qualification calls and all proposal attempts; raw-request costs total \$0.24462614. No local or rented GPU was provisioned. The fixed 40-start design was not enlarged after the unresolved assignment contrast. Outcome availability transports under this contract; full correct-assignment dissociation remains unresolved. Shared biology, laboratories, exposed-source selection and unverified hosted weight identities continue to bound transfer.

M Statistical definitions and policy-prediction details

The following definitions retain the complete estimator and resampling specifications. The compact symbols in numerical tables are decoded here; all effects concern complete recovery-inclusive policies on fixed biological sources.

We distinguish two questions: does the LLM benefit from feedback, and does it beat a simpler policy with the same access? Let $Y_{p,a,e}$ be final cumulative hits for controller p , feedback a and executor e . In the equations, L/U denote LLM/uniform-hit; $T/W/O$ denote TRUE/WITHIN/NO (correct, permuted or withheld outcomes); and N/R denote NN/random expansion. The LLM’s correct-assignment and outcome-availability benefits are $B_{\text{assign}} = \mathbb{E}[Y_{L,T,N} - Y_{L,W,N}]$ and $B_{\text{avail}} = \mathbb{E}[Y_{L,T,N} - Y_{L,O,N}]$. Permuting outcomes can actively mislead a policy, whereas withholding them removes information. The two benefits are therefore different. For example, a positive assignment benefit can coexist with negative controller advantage if the simpler rival benefits even more from correct feedback. Four primary B640 contrasts compare controllers, their feedback benefits and their dependence on NN expansion:

$$\begin{aligned}\Delta &= Y_{L,T,N} - Y_{U,T,N}, \\ \Gamma &= (Y_{L,T,N} - Y_{L,W,N}) - (Y_{U,T,N} - Y_{U,W,N}), \\ \Omega_L &= Y_{L,T,N} - Y_{L,T,R}, \quad \Omega_U = Y_{U,T,N} - Y_{U,T,R}.\end{aligned}$$

Δ is LLM advantage over uniform-hit; Γ compares their benefits from correct outcome assignment; each Ω measures the benefit of NN rather than random expansion. Dissociation describes positive feedback benefit together with negative advantage; it adds no endpoint. Controller-by-executor interaction is $\Omega_L - \Omega_U$. Paired-start effects equally weight 36 B640 model-screen cells; twenty starts/cell are resampled 20,000 times [Efron, 1979, Agarwal et al., 2021]. Four primary 98.75% percentile intervals use Bonferroni family adjustment. These intervals quantify retrospective contrasts, not a deployment decision threshold. Other balanced 95% intervals are supporting, including the entire B160 panel.

The external controller study reports LLM-minus-uniform-hit with primary 95% model-stratified intervals. The external feedback study tests four contrasts: controller advantage, correct-assignment benefit, outcome-availability benefit and the difference in controllers’ assignment benefits. It uses 98.75% intervals from 20,000 paired bootstrap draws within model, with equal model weights.

Each follow-up study defines its own family of primary comparisons. Adjusted intervals address that family’s multiple comparisons; supporting 95% intervals answer individual descriptive questions.

The deployment study has five primary comparisons: LLM advantage (Δ), the difference in controllers' assignment benefits (Γ), each controller's benefit from NN expansion (Ω_L and Ω_U), and the LLM's assignment benefit (B_{assign}). Its 99% intervals equally weight the two sources and the Gemini/DeepSeek recipes; starts are resampled within source jointly across policies and models. Qwen comparisons are supporting analyses, not members of this pool.

The separate Gemini/DeepSeek initializer-controller study uses a twelve-test family. Its 99.583% intervals come from 50,000 paired bootstrap draws, resampling starts within source jointly across conditions/models, with equal source weights. Source intervals and other contrasts are supporting 95%. The fixed sample size guarantees neither power nor equivalence.

Sensitivity checks vary hit criteria ($|z| \geq 2$), score standardization, round, source and proposal completeness. Standardization uses evaluation-only whole-screen moments; costs include failures. Intervals condition on fixed sources [Dror et al., 2018, Demsar, 2006]; no significance-driven sampling or post-hoc equivalence margins were used [Lakens, 2017].

To test whether feedback benefit and controller advantage coexist on individual sources, we jointly resample their contrasts (Figure 2C). Equal-cell weighting gives each model-readout combination the same weight; equal-source weighting prevents sources with two readouts from counting twice.

The full source-and-family-excluded prediction protocol, fixed penalty, training-only preprocessing and permutation test are given with their results in Appendix L.

M.1 Interface and selection definitions

The naming and response-budget protocols, eligible centre menus and interval families are specified in Appendix H. Both are first-adaptive tests on 840 original-panel states, with 720 B128 states primary and 120 B32 states supporting. They retain shared recovery and do not establish full-campaign effects.

We count hits at proposed centres separately from hits among their neighbours. At these same 840 first-adaptive states, we also rerun NN with the proposed centres retained as anchors but excluded from acquisition. In centre order, each anchor fills its original quota with cosine-nearest untested genes not already selected. This recomputes the entire batch at the same budget.

A geometry-based selector chooses between LLM and uniform-hit proposals after NN expansion. For each acquired gene that is not itself a proposed centre, we take its cosine distance to the centre that selected it, then average equally over those genes. The selector prefers the smaller mean, breaking ties in favor of uniform-hit. Thus it measures how close the acquired neighbours are to their anchors, rather than all pairwise distances within the batch.

For Llama and Qwen2.5, we make new model calls along each policy's own history, using ten previously used starts per model-screen cell across seven screens. Two 97.5% B640 contrasts compare this selector with uniform-hit and with the same selection rule given two uniform-hit proposals. This second rival tests whether improvement comes from choosing between two proposals at all, rather than from including an LLM proposal.

Terms and table labels

Term	Meaning in this paper
Initializer	Chooses the first acquired batch; those tests count toward the total budget.
Controller / centres	Proposes five gene anchors for subsequent acquisition. An anchor may be previously tested, while acquired genes must be new.
Executor / NN	Expands centres into a fixed-size batch; NN uses cosine neighbours in normalized Achilles features.
Recovery	Deterministic, outcome-blind completion of missing legal centres. It is included in pipeline performance, not credited as model generation.
Uniform-hit / first-hit	Selects displayed hits randomly / in recorded acquisition order. Both use the same executor; under NO they cannot filter by hit status.
TRUE / WITHIN / NO	Correct gene-outcome pairs / intact within-batch pair reassignment / withheld outcomes with tested identities retained.

Term	Meaning in this paper
Random expansion	Retains the centres that NN would acquire at that state and fills other slots uniformly. It differs from wholly random acquisition.
Δ, Γ	LLM-minus-uniform-hit yield; LLM-minus-uniform difference in correct-assignment benefits. Negative values favor the simple rival on the corresponding comparison.
Ω_L, Ω_U	NN-minus-random expansion benefit for the LLM and uniform-hit controllers. Their difference measures executor dependence of controller advantage.
$B_{\text{assign}}, B_{\text{avail}}$	Within-LLM correct-minus-permuted and correct-minus-withheld yield. They measure information benefit, not superiority over another controller.
Start / source / read-out	A computational randomization / an outcome-source lineage / a measured endpoint. More starts or two readouts from one source do not create biological independence.
B640 / B160	Five-round budgets of 640 / 160 distinct gene tests. Their source populations differ.