
Position: Toward a Computational Theory of Scientific Discovery

Odysseus Kaloidas
UC Berkeley

Abstract

1 As AI improves, which scientific problems become solvable next, and when? In
2 a hypothetical assay comparison, the first procedure to become available can be
3 less likely to deliver a result by a common deadline. We argue that a computational
4 theory of scientific discovery should therefore connect capability growth to target
5 success, procedure availability, and validated completion. Two retrospective studies
6 test whether information about the target improves forecasts. Across 73 models
7 and two reasoning families, object count improves pooled prediction but worsens
8 prediction of larger deduction tasks; it does not isolate computational complexity.
9 In 550 chemical-table replays, early outcome summaries worsen forecasts for
10 the original easy hit criterion. They help with a stricter criterion at the shorter
11 observation budget, but that advantage reverses at the longer budget. Fitting each
12 budget separately does not remove that reversal. Agreement with the overall
13 success rate also conceals opposing errors across chemical sources. Conditional
14 calculations explain which measurements constrain availability dates and why
15 validation and evaluation costs matter. Together, these comparisons motivate
16 a research programme for choosing informative measurements and comparing
17 research procedures; prospective completion dates remain unvalidated.

18 1 Introduction

19 As AI progresses from olympiad problems to research proofs, what changes in our expectations about
20 an open conjecture, an unsolved algorithmic problem, or a cancer treatment? Predicting what comes
21 next requires knowing which target requirements a system can meet and what validation remains.
22 This question arises within mathematics as well as between mathematics and biomedicine.

23 Consider a procedure that selects a compound and tests it in a specified cell assay. A mathematical
24 breakthrough alone does not tell us how reliably this procedure will find a confirmed hit. Even if that
25 reliability can be predicted, two further questions remain: when will a suitable procedure become
26 available, and when will selection and validation finish? The procedure available first need not be
27 most likely to finish by a deadline (§5). This comparison is hypothetical; patient benefit requires a
28 different target.

29 Our position is that a computational theory of scientific discovery should connect capability growth to
30 the changing set of achievable targets and to their eventual completion. We first need a target response:
31 how measured capability relates to the chance of validated success. Combining that response with a
32 scenario for future capabilities and resources gives an availability date. Deployment and validation
33 then determine completion. A breakthrough can inform the target response without settling these
34 later stages. Researchers need target measurements before extrapolation and validation schedules
35 before choosing a procedure. An unknown date does not mean the target is unsolvable.

36 The assay example supplies exact calculations of measurement and evaluation costs (§§4–6). Two
 37 archive studies ask whether task structure or early experimental outcomes improve prediction (§§3,7).
 38 Their gains and reversals show why added information must be tested on the intended targets. These
 39 conditional examples and retrospective tests do not validate a calendar of breakthroughs.

40 These foundations are established: complexity and learning theory separate computation from
 41 information requirements [Arora and Barak, 2007, Shalev-Shwartz and Ben-David, 2014, Traub et al.,
 42 1988]; observational scaling predicts downstream performance from capability coordinates [Ruan
 43 et al., 2024]; Sloth connects latent skills to model size and training tokens [Polo et al., 2025]. ADeLe
 44 uses annotated task demands [Zhou et al., 2026]; other predictors use performance histories and
 45 validation loss [Hu et al., 2026, Truong et al., 2026]. Capability theories, measurement validity and
 46 biomedical acceleration constraints also precede this work [Jo and Wilson, 2026, Wallach et al., 2025,
 47 Hebenstreit et al., 2026]. Our synthesis connects these established tools to the choice of scientific
 48 procedure and tests where extra target information helps. It introduces no new scaling law: a response
 49 model combined with the correct deployment model already implements this position.

50 2 Defining the target and its conditions

51 2.1 Define the scientific event and its resources

52 A target Q specifies the problem or instance population and the validated outcome to be produced.
 53 Its success probability can refer to two settings. For a declared distribution of problem instances,
 54 success is averaged over those instances. For one fixed problem, the probability instead concerns
 55 randomness in the procedure and its observations.

56 For the assay, fix a cell line, viability threshold and repeat-assay confirmation rule before selecting
 57 compounds. The target is one compound meeting that rule; an attempt includes selection and
 58 validation. A correct candidate somewhere in a generated set does not establish that the procedure can
 59 select and validate it [Dorner et al., 2026]. A causal mechanism or patient benefit would be a different
 60 target. Likewise, proving a supplied statement differs from selecting and resolving a significant open
 61 conjecture.

62 To retain trade-offs between computation, observations and time, we record resource combinations.
 63 Let K denote information and observation access, Π the permitted procedures, and $\mathbf{b}$ the resource
 64 allowance. A procedure includes its model, tools, orchestration and human assistance. Comparing
 65 complete systems permits these to vary; attributing a gain to one component requires controlling the
 66 others. For procedure π , write $p(\pi, Q; \mathbf{b}, K)$ for its probability of producing a validated result within
 67 those limits. At required reliability q , define

$$\mathcal{B}_Q(q \mid K, \Pi) = \{\mathbf{b} : \exists \pi \in \Pi, p(\pi, Q; \mathbf{b}, K) \geq q\}, \quad \mathcal{A}_t(q) = \{Q : \mathbf{b}(t) \in \mathcal{B}_Q(q \mid K_t, \Pi_t)\}. \quad (1)$$

68 The first set contains sufficient resource combinations; the second gives achievable targets under
 69 resources $\mathbf{b}(t)$, information K_t and procedures Π_t available at time t . In the assay example, computa-
 70 tion, allowed assays and elapsed validation time are separate resources. More computation may help
 71 select candidates without reducing the duration of their assays. The allowance includes campaign
 72 duration, so achievable at time t means a suitable campaign can begin then, not that it has already
 73 finished.

74 If two procedures need (10 compute units, 10 observations) and (100, 1), their separate minima do
 75 not imply that (10, 1) suffices. Existence of a suitable procedure does not mean an evaluator finds
 76 it. Inherent computational complexity instead concerns all algorithms in a specified model, often
 77 asymptotically. Today’s model failure establishes neither an inherent lower bound nor a date for
 78 overcoming it.

79 Required reliability q is a decision requirement, not a complexity estimate. Its choice must reflect
 80 the consequences of failure. Mean quality cannot replace success probability, which depends on the
 81 validation and selection procedures and their errors.

82 2.2 Observation access changes the attainable set

83 Some targets require information that the allowed observations cannot supply. Suppose two mecha-
 84 nisms generate identical distributions of observation histories under every permitted adaptive policy,

85 but require opposite binary answers. Every permitted procedure then has the same output distribution
86 in both worlds: choosing one answer with probability r gives accuracies r and $1 - r$. Worst-case
87 accuracy is at most $1/2$. This premise constrains every procedure; poor performance by one predictor
88 does not establish it. Distinguishing observations or a justified structural restriction can remove this
89 obstruction; more computation on unchanged evidence cannot. This is an identification argument
90 under explicit observation restrictions, with close connections to causal transport and biological
91 context [Bareinboim and Pearl, 2016, Dibaenia et al., 2026, McCormick, 2026].

92 For example, recover the program $f_a(x) = ax \bmod 2^n$ with uniformly hidden odd a . If each
93 rejected candidate only rules out its own coefficient, positive verification with 90% probability needs
94 $\lceil .9 \cdot 2^{n-1} \rceil$ checks. Observing $f_a(1)$ instead reveals a . The exponential cost belongs to the restricted
95 interface, not multiplication. Appendix C supplies exhaustive checks and verification costs.

96 3 From demonstrated capability to target performance

97 A source score describes a system; it does not describe what a new target demands. We ask whether
98 combining the two improves prediction. A solved problem can reveal relevant ability without
99 transferring its solution. Let $\mathbf{z}$ be the measured capability profile and $\kappa_Q(K)$ the target description
100 under access K . A shared rule F_θ , with fitted parameters θ , predicts success from these quantities and
101 the resource allowance. For the assay target, success means one complete selection-and-validation
102 attempt. Our archive tests instead use computational outcomes. The second expression is a restricted
103 logistic example:

$$\hat{p}_Q = F_\theta(\mathbf{z}, \kappa_Q(K), \mathbf{b}), \quad \hat{p}_Q = \sigma(\mathbf{a}_Q^\top \mathbf{z} - d_Q), \quad \sigma(u) = (1 + e^{-u})^{-1}. \quad (2)$$

104 The logistic model holds resources, information and protocol fixed. Log-odds are $\log[p/(1-p)]$; a
105 linear relation on this scale becomes a curved probability response. The weights $\mathbf{a}_Q$ describe how
106 target log-odds vary with measured capabilities; d_Q shifts the response curve. This item-difficulty
107 parameter is not computational complexity. Fix capability units and normalization before comparing
108 coefficients.

109 For an unseen target, shared functions must determine its coefficients: $\mathbf{a}_Q = \mathbf{a}_\theta(\kappa_Q)$ and $d_Q =$
110 $d_\theta(\kappa_Q)$. Fitting separate coefficients after observing each target’s outcomes cannot supply that
111 prediction. Even a successful shared fit remains an association: it does not establish that intervening
112 on one capability causes improvement.

113 3.1 Testing target descriptions in a common archive

114 We use the public ObsScaling archive [Ruan et al., 2024]. Principal component analysis (PCA)
115 summarizes seven source scores: MMLU, ARC-C, HellaSwag, Winograd, TruthfulQA, GSM8K and
116 XWinograd. PC1 is the first coordinate; PC1–3 uses the first three. Source scores and PC coordinates
117 are standardized on training models only.

118 We test a narrower version of the shared-description hypothesis using BIG-Bench Hard (BBH) logical
119 deduction and shuffled-object tracking, each with 3, 5 and 7 objects. The descriptor is log object
120 count, fixed from task definitions. Three-shot chain-of-thought strict-match scores come from the
121 same archive: 73 models in 20 archive-defined groups, each evaluated on six task versions, giving
122 438 aggregate success rates. These 20 archive groups can share ancestry; their identities are listed
123 in Appendix A. A cell is a published model–task rate, not one generated answer. We compare PC1
124 and PC1–3 with count-informed predictors. Adding count shifts predicted log-odds; interactions
125 additionally let capability effects vary with count.

126 PCA uses each training model once; regressions fit its rates on the available task versions. Prediction
127 inputs are seven measured source scores and, for count-informed fits, target object count. Family
128 labels define exclusions; only the task-family-mean baseline uses task identity as a prediction key.
129 Test rates are withheld. We score predictions by mean squared error against archived success fractions
130 (rate-MSE), not by counting validated discoveries. Table 1 separates task-family, model-family and
131 size holdouts from joint family exclusion.

132 In Table 1, PC1–3 + count outperforms the interaction model in the three separate holdouts. The
133 chance-floor predictor uses $1/n + (1 - 1/n)\sigma(\beta_0 + \beta^\top \mathbf{z})$ for n objects; it assumes rather than
134 guarantees a floor on LLM accuracy. The task-family mean uses training rates from the same task

Table 1: Archived rate-MSE; lower is better. Separate holdouts exclude a task family (438 cells), model family (438), or size 7 (146; training at 3/5). Joint holdout excludes both families (438 cells); Cells weights cells equally, Families weights the 20 model groups equally. Each split uses identical cells across predictors. † marks post-result rivals; the joint test is also post-result. These are retrospective, not contamination-controlled tests.

Predictor	Separate holdouts			Joint holdout	
	Task	Model	Size	Cells	Families
Pooled mean	0.0431	0.0427	0.0446	0.0446	0.0532
Task-family mean	0.0431	0.0422	0.0447	0.0446	0.0532
PC1	0.0342	0.0320	0.0307	0.0357	0.0432
PC1-3	0.0355	0.0330	0.0307	0.0379	0.0459
Count only†	0.0395	0.0371	0.0279	0.0410	0.0535
PC1-3 + count†	0.0260	0.0234	0.0179	0.0285	0.0356
PC1-3 + chance floor†	0.0266	0.0270	0.0217	0.0287	0.0350
PC1-3 + count interactions	0.0277	0.0238	0.0195	0.0301	0.0377

family, falling back to the pooled mean when that family is withheld. Among these eight, the chance-floor predictor has the lowest task-family cross-entropy (Appendix A); no new models were evaluated.

One held-out prediction. To predict Gemma-7B’s seven-object deduction score, we withhold its model group, the deduction family and size 7, then train on 142 tracking rates: 71 non-Gemma models from 19 groups, at 3/5 objects. Gemma-7B’s source scores and the target count enter the fitted rule; its deduction outcome does not. PC1-3 predicts .411; adding count predicts .211. The archived rate is .376, giving squared errors .00120 and .02720. This example shows how one prediction is scored; it does not represent the average effect. Figure 1 summarizes all such exclusions, including the pooled benefit and deduction reversal.

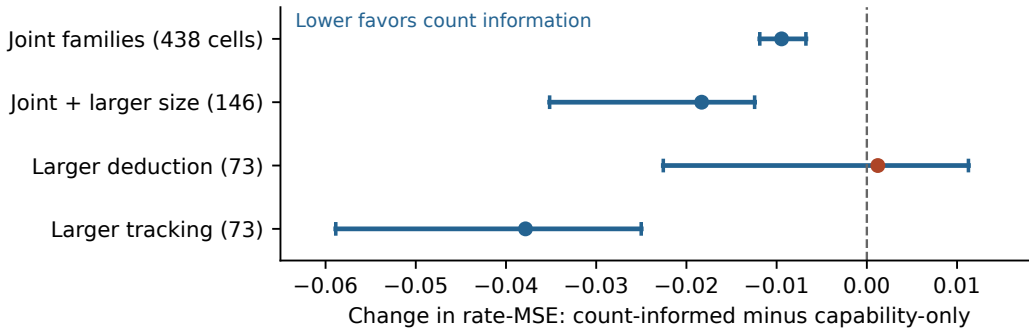


Figure 1: Does object count improve prediction when both model and task families are withheld? Dots show the rate-MSE of PC1-3 + count minus that of PC1-3; negative favors adding count. Bars are 95% percentile ranges from 500 model-family refits, not uncertainty over new task families. The last two rows partition the 146-cell size-extrapolation test. Deduction worsens at the observed point estimate and its interval spans zero.

PC/count fits minimize the sum of squared residuals on logits of target rates clipped to [.005, .995], plus the sum of squared non-intercept coefficients (ridge penalty 1). Log count is standardized on training cells. Coefficients are shared across task versions; the intercept is unpenalized. The penalty is fixed, not tuned on held-out outcomes.

At seven objects, adding count changes tracking rate-MSE from .06159 to .02375, but deduction from .01680 to .01802 (73 cells each; Figure 1). This establishes a target-description association, not a computational-complexity mechanism: count also changes the answer space. A forecast for a new target needs a relevant transfer test, not just a better pooled score.

The preferred predictor also depends on whom the average represents. In Table 1, additive count narrowly beats the chance floor when cells receive equal weight (.02845 vs. .02873), but loses when each model family gets weight 1/20 (.03558 vs. .03503). Rankings therefore depend on the error measure, weighting and holdout. The pooled count gain survives alternative fits (probit, direct rate prediction, clipping thresholds and a quadratic capability-only rival), but adding count to the chance-floor predictor is not uniformly better.

Figure 1's ranges concern model-family sensitivity with two fixed task families and no item-level outcomes (Appendix A.4). A date forecast additionally needs a capability-growth trajectory.

4 From target performance to a conditional date

4.1 The missing relation and how to measure it

A response curve gives success at a capability level; a date asks when that procedure becomes available. For the assay, require a complete single-candidate procedure to pass with probability at least q , rather than reaching q through many attempts. For fixed reliability and a development scenario, define the earliest availability time as

$$\tau_Q(q) = \inf\{t \geq 0 : Q \in \mathcal{A}_t(q)\}. \quad (3)$$

If no permitted procedure ever meets the target, $\tau_Q = \infty$. This date concerns availability; a campaign need not start then. Each scenario below permits one procedure; admitting alternatives cannot delay the earliest availability date. A probabilistic forecast would also need uncertainty about capability growth, target response, resources and access.

The following values are hypothetical. Here p_Q is single-attempt success and q its required threshold; λ measures log-odds change per capability unit, v capability units per year, and τ_Q years to availability. Section 6 instead uses ρ for a multi-attempt campaign hit probability. With initial log-odds η_0 and a fixed linear relation,

$$\text{logit } p_Q(t) = \eta_0 + \lambda vt, \quad \tau_Q(q) = \frac{\text{logit } q - \eta_0}{\lambda v}. \quad (4)$$

The finite crossing formula assumes $v, \lambda > 0$ and initial success below q . With $\eta_0 = -3$, $v = 1$ and $q = .8$, procedures A and B with slopes (couplings) $\lambda = 1$ and $.5$ become available after 4.39 and 8.77 scenario-years; a zero-coupling procedure never does. All three agree on the source trajectory and current target success .0474. At one additional source unit, A and B instead predict .1192 and .0759: a matched target measurement can begin to distinguish them, subject to sampling error. Thus source progress and current target performance do not identify the horizon when coupling is unrestricted. Appendix Figure 2 displays these constructed curves.

A constructive measurement procedure. Measuring the target at two capability levels supplies information missing from a single current success rate. Assume log-odds vary linearly with capability. Use the same target protocol at levels separated by $\Delta > 0$, with N independent success/failure trials at each level. Form individual 97.5% Clopper–Pearson intervals $[L_0, U_0]$ and $[L_1, U_1]$ by allocating .0125 to each of four tails, giving at least 95% simultaneous coverage [Clopper and Pearson, 1934].

Subscripts 0 and 1 denote lower and higher capability. To bound improvement conservatively, pair the lower success bound at higher capability (L_1) with the upper bound at lower capability (U_0); reverse the endpoints for the largest change. On the log-odds scale, dividing by the capability separation gives

$$\lambda \in \left[\frac{\text{logit } L_1 - \text{logit } U_0}{\Delta}, \frac{\text{logit } U_1 - \text{logit } L_0}{\Delta} \right]. \quad (5)$$

Across repeated samples at these fixed levels, at least 95% of intervals cover the slope if the binomial and affine-logit premises hold; this is not 95% coverage of a future discovery date. Zero successes or all successes at a level can leave an interval endpoint infinite. Cross-level independence is unnecessary for coverage, but used for exact power calculations. The slope describes association unless capability is isolated by intervention. At observation cost a , measurement costs $2Na$.

We enumerate success-count pairs at sample sizes 25/100/500, initial yields .01/.05/.2, separations .5/1/2 and couplings 0/.5/1. These 81 prespecified settings check coverage and recovery, including

zero coupling and low yield. With initial target success .05, true coupling 1 and separation 2, the probability of a positive lower coupling bound is 4.9% at $N = 25$, 90.7% at $N = 100$, and above 99.9% at $N = 500$ per level. At $N = 100$, the median interval width is 1.69 and 0.59% of intervals remain unbounded. The finite grid’s minimum coverage is 99.81%; the analytical union-bound argument supplies the general guarantee (Appendix B).

At the same $N = 100$, slope 1 and separation 2, reducing initial success to .01 lowers positive-slope detection to 6.585% and leaves 36.651% of intervals unbounded. The measurement can therefore be valid yet uninformative. In the assay example, very few validated hits at either capability level would leave the improvement rate poorly constrained even when both groups were measured correctly.

Bounding the availability date requires a positive lower slope bound, a supported growth path and continued access. Fix the endpoint, protocol and source scale before sampling; propagate joint uncertainty to the target threshold. A rising but saturating response can match both probabilities exactly and never reach the threshold (Appendix B). Fit on two levels, then predict a withheld third before observing it; repeat across held-out targets. This tests extrapolation beyond the fitted pair. Logistic is a starting link, not a privileged law; archive link sensitivities are retained.

Repeated trials estimate success on the declared target; they do not sample new problems. Model-family comparisons may change data, architecture and tools together. Binomial intervals do not remove these confounders: record the complete procedure and check its stability across levels.

4.2 Why accurate performance fits can give poor dates

Near a plateau, a small error in predicted success can imply a large error in the date it reaches a threshold. The response slope determines that conversion. If true target log-odds $\ell(t)$ have derivative at least $m > 0$ throughout an interval containing both the true and predicted crossings, and predicted log-odds have uniform error at most ε there, then

$$|\hat{\tau} - \tau| \leq \varepsilon/m. \quad (6)$$

At the predicted crossing, the true log-odds are at most ε from the threshold. A slope of at least m forces that gap to close within ε/m time units (Appendix D.1). A log-odds error of .2 thus allows .2 years at slope 1 per year, but 2 years at slope .1. The difficulty is establishing uniform error and positive growth over the relevant future interval. Pointwise fitted-curve intervals do not establish either premise.

5 From availability to validated completion

Return to the assay target. A confirmed hit arrives only after a campaign is launched and validation finishes; procedure availability alone does not specify completion. Let C_Q be that completion time. For a fixed schedule with validation-completion dates t_j , let h_j be the probability that campaign j succeeds conditional on all preceding campaigns failing. The probability of no success is the product of these conditional failure probabilities, so

$$\Pr(C_Q \leq T) = 1 - \prod_{j:t_j \leq T} (1 - h_j). \quad (7)$$

This probability identity is not a fitted scaling law: prediction still requires the conditional success probabilities under the deployment policy. Marginal scores cannot generally replace them: earlier failures may reveal a shared weakness that makes later campaigns less likely to succeed. An attempt may contain an adaptive agent workflow; independent restarts are a separate assumption. With independent serial attempts of success p , launch time s and validation duration d , the expected number of attempts is $1/p$, giving mean completion $s + d/p$.

One decision carried from capability to completion. Continue the hypothetical assay comparison in Section 4. A reaches reliability .8 at year 4.39; B at 8.77. Choose one procedure, launch immediately on availability, and freeze that version at success .8. Suppose independent serial attempts require four years of validation for A and one quarter-year for B. One validation pipeline is available; further model improvements during a campaign are excluded. By a common deadline at year 10, A completes one attempt and B completes four, giving success probabilities .8 and $1 - (1 - .8)^4 = .9984$.

244 Their mean completion times are 9.39 and 9.09 years. Choosing the earlier available procedure
245 therefore chooses the worse completion forecast in this scenario.

246 At the same fixed success q , B has lower mean completion time precisely when $d_A - d_B > q(\tau_B - \tau_A)$.
247 Here the duration advantage is 3.75 years, exceeding the 3.51-year boundary. Measuring validation
248 duration can change the choice even after capability responses are known. If durations are equal, A
249 wins; if access or attempt dependence changes, recompute the comparison. Combining the success
250 forecast with validation duration already gives this answer; the example shows why both are needed.
251 All values are stipulated, not biomedical forecasts.

252 For a fixed agent computation, let W be total work and L the duration of the longest chain of dependent
253 steps. With P processors, runtime satisfies $T_P \geq \max\{W/P, L\}$. If validation requires r successive
254 rounds of duration at least ℓ , then $L \geq r\ell$. More agents can change the procedure; parallelizing
255 unchanged work preserves its dependency chain [Meyerson and Qiu, 2025, Shun, 2024, Kim et al.,
256 2026]. Agent-count comparisons must include context processing, coordination and validation costs.
257 Serial-computation bounds depend on architecture [Liu et al., 2026b]; human-duration horizons differ
258 from workflow depth [Liu et al., 2026a].

259 6 Scientific targets and the cost of establishing progress

260 6.1 What the scientific example measures

261 Co-Scientist’s initial acute myeloid leukemia (AML) repurposing example has three cell-viability hits
262 among five oncologist-selected drugs from thirty ranked proposals [Gottweis et al., 2026]. This is a
263 selected-subset result for a human-assisted procedure. These selected assay hits support that source’s
264 demonstration, but do not estimate a capability–yield curve: capability levels were not paired with
265 matched laboratory outcomes. Clinical benefit requires different validation (Appendix Table 10).

266 6.2 A worked decision about future campaigns

267 Establishing a procedure’s advantage may itself be expensive. The following calculation separates
268 low decision error from useful, affordable evaluation. This hypothetical pilot evaluates method
269 selection: its attempts are charged separately and their own discoveries receive no value. Here p is
270 per-attempt success; ρ is the required chance of at least one campaign hit, distinct from single-attempt
271 reliability q in Section 4. Independent attempts need a cap of $\lceil \log(1 - \rho) / \log(1 - p) \rceil$. At $p = .1$
272 and $\rho = .9$, this is $\lceil \log(.1) / \log(.9) \rceil = 22$ attempts. At costs of 10 and 11 units per attempt, the
273 baseline and candidate need reserves of 220 and 242: enough budget to fund their attempt caps.
274 Doubling candidate yield to .2 reduces its cap to 11 and reserve to 121; a strict saving already occurs
275 at approximately .1141. The pilot estimates both yields. Reserve differs from expected expenditure
276 until first success [Erol et al., 2026]: the baseline averages $10/.1 = 100$ units with recognizable
277 success and early stopping, although its 90%-success cap is 220.

278 Take 100 independent success/failure observations per arm, with independent arms. Individual 97.5%
279 Clopper–Pearson yield intervals (tail probability .0125) give at least 95% simultaneous coverage.
280 Convert each yield interval to reserve bounds using the cap formula: lower yield needs more attempts,
281 so it gives the upper reserve. The rule adopts only when savings hold throughout these bounds. It
282 adopts if the candidate’s upper reserve is below the baseline’s lower reserve, declares non-saving if
283 the candidate’s lower reserve is at least the baseline’s upper reserve, and otherwise defers. Deferral
284 means insufficient evidence for either decision, not that the candidate cannot save resources. Zero
285 lower yield gives an infinite upper reserve.

286 For example, ten baseline successes and forty candidate successes give reserve ranges [120, 520] and
287 [44, 77] units. Even the candidate’s upper bound, 77, is below the baseline’s lower bound, 120: the
288 rule adopts. Ten successes in each arm give overlapping ranges [120, 520] and [132, 572], so it defers.

289 Enumerating every pair of success counts, weighted by its probability under the stipulated yields,
290 gives exact decision frequencies rather than empirical estimates. At equal yields .1, comparing
291 reserves at the observed success fractions falsely favours the candidate in 41.877% of pilots. The
292 interval rule falsely adopts in .018% but leaves 99.900% unresolved. At candidate yield .2, it adopts in
293 only 4.572%. Among resolved decisions, the error rate is 18.33% at unchanged yield: unconditional
294 error control does not imply 95% correctness conditional on deciding. Always deferring gives zero

295 false adoptions but also zero useful adoptions, without pilot cost. The interval rule must earn its cost
296 through correct decisions, not low error alone.

297 At $N = 1000$ and candidate yield .20, adoption rises to 98.503% (Appendix Table 11), but the
298 evaluation costs 21,000 pilot units for a true reserve saving of 99 per future campaign. Even granting
299 perfect adoption, offsetting that cost takes 213 campaigns; this is reserve accounting, not cash
300 payback. Appendix E.4 reports threshold sensitivities. The illustrative rule is not optimized and
301 establishes no intrinsic evidence-cost limit.

302 7 Early experimental outcomes and forecast limits

303 7.1 Do early experimental outcomes improve a completion forecast?

304 We replayed 550 campaigns: 50 seeds on each of eleven 512-candidate tables from four source
305 groups: four additive-screening plates, five Buchwald–Hartwig tables, one C2–yield table and one
306 Suzuki–Miyaura table [Ranković et al., 2026]. Each campaign starts with five random observations,
307 then selects the candidate maximizing kernel-predicted yield plus one posterior standard deviation.
308 All candidate input features are known. The kernel is $\exp(-2d)$, where d averages numeric squared
309 distances normalized by catalogue ranges and categorical mismatches; diagonal regularization is .01.
310 Each selection reveals one recorded outcome. Seeds vary histories on fixed tables, not independent
311 scientific validations.

312 After five or ten observations (the pilot), we predict whether that campaign will find a hit by
313 observation 20 or 40 (the total budget). All forecasts receive these counts and whether a hit has
314 already occurred. The hit threshold is the complete table’s 90th percentile: it uses unrevealed
315 outcomes and is supplied retrospectively, not estimated from the pilot. Apart from this threshold,
316 unrevealed yields are not predictor inputs. After a pilot hit all methods assign probability one.

317 Otherwise, the budget-only logistic regression uses log pilot length and log total budget; the richer
318 predictor adds threshold-relative best and mean yields, yield dispersion and mean pairwise kernel
319 similarity. We withhold one entire source group at a time. Fits use only no-hit pilots from the other
320 three groups, train-only feature standardization, non-intercept ridge penalty 1, and weights balancing
321 tables and groups. Thus the comparison tests whether early outcome summaries improve forecasts
322 beyond campaign size, not whether more capable AI discovers better chemistry (Appendix G).

323 Brier score measures squared probability error against the campaign’s binary outcome, averaging
324 tables within source groups and then groups equally. After five observations, forecasting a hit by
325 observation 20 gives .01862 with summaries versus .01773 for budget-only (lower is better). Among
326 329 campaigns without a pilot hit, scores are .03253 versus .03106. Neither comparison favors the
327 richer predictor. The group-averaged hit rate reaches 98.2% by budget 20; every retained campaign
328 hits by 40. Later checks give Brier scores .01774 for the success rate estimated from training groups
329 (prevalence) and .01800 for always predicting success. Always-success is worse on log-loss. This
330 endpoint limits discrimination; the replay establishes no wet-lab discovery.

331 **Does a stricter endpoint change the conclusion?** After observing near-universal success, we
332 compared the original 90th-percentile endpoint with stricter 95th and 99th percentiles, keeping the
333 same trajectories and source exclusions. Changing the endpoint changes hit labels, pilot-hit status
334 and threshold-relative features, while the search policy stays fixed. At the 99th percentile and budget
335 20, summaries outperform budget-only, best-gap and training prevalence (Brier score .28510). Their
336 Brier improvements over budget-only are .17351 (additives), .06007 (Buchwald–Hartwig), .11672
337 (C2) and .13881 (Suzuki–Miyaura). At the original and 95th-percentile endpoints, summaries worsen
338 all four groups. The strict-endpoint gain reverses at budget 40 (Table 2). Thus the original non-win
339 does not establish that early observations are uninformative, and the later gain does not establish
340 uniformly better forecasts. This is retrospective sensitivity, not independent confirmation.

341 **Is the reversal just an overly restrictive forecaster?** The original regression shares outcome-
342 summary effects across budgets. Separate budget-specific fits test this restriction using the same
343 histories, source exclusions and training-feature scales. At the 99th percentile after five observations,
344 the longer-budget summary score worsens from .29591 to .33922, versus .22607 for budget-only.
345 Separate fitting therefore does not remove this reversal. Appendix G.3 gives the penalty control and
346 all settings, including forecasts that incorrectly assign less chance of a hit with more observations.

Table 2: Same 550 histories, all three hindsight endpoints and both forecast budgets after five pilot observations. Brier scores: lower is better; hit rates and errors weight tables within sources, then the four sources equally. Best gap adds the best observed threshold-relative outcome to budget-only. The 99th-percentile gain reverses at budget 40.

Percentile	Budget	Hit rate	Budget-only	Best gap	Summaries
90	20	98.2%	0.01773	0.01826	0.01862
90	40	100.0%	0.00000	0.00000	0.00000
95	20	89.4%	0.09346	0.10043	0.10460
95	40	99.5%	0.00503	0.00514	0.00515
99	20	42.0%	0.28716	0.34035	0.16489
99	40	80.9%	0.22316	0.22688	0.29591

The overall success-rate average also hides opposing source errors. At the same longer budget, original summary forecasts average 81.9% against 80.9% observed. Yet additives are underpredicted (35.0% versus 100%) and C2 overpredicted (96.9% versus 28%). These are descriptive held-out errors, not new independent experiments. Agreement in the overall mean, even with probabilities increasing with budget, is insufficient: a discovery forecast must also be tested on the target populations where it will guide research.

7.2 Alternative views and limits

A small capability representation with target calibration may suffice when access, tools and validation stay stable. Extra descriptors can be costly, noisy or redundant, as our deduction reversal and easier chemical endpoints illustrate. The question is which omissions change a decision, not how elaborate a predictor is. Scientific progress needs no computational theory; the theory earns its value through better comparisons, forecasts or measurement choices.

Representations and instruments may change target response and access [Ge et al., 2026, Lourie et al., 2025, Hernandez et al., 2021, Longpre et al., 2026]. Neither retrospective study tests that change causally. The descriptor study has only two reasoning families, shared training lineages and benchmark items, no item-level uncertainty, and uncontrolled contamination. The chemistry study has four heterogeneous origins, hindsight thresholds and fixed archived outcomes. More random seeds do not create new sources; their loss differences do not establish a population confidence interval. Neither study tests whether this framework improves a real research allocation decision or predicts a prospective scientific completion date.

For an independent test, fix the endpoint and input access before observing new targets. Fit at two capability levels and predict a third against capability-only, target-only, no-change and response-plus-duration forecasts. Measure costs, validation delays and censored non-completions [Gneiting and Raftery, 2007]; AstaBench offers infrastructure [Bragg et al., 2026]. Separating answer choices from count, and work from depth, is additionally necessary for complexity claims. Retain simpler forecasts when added information fails to improve decisions after acquisition cost. Assay activity establishes neither clinical safety nor permission for autonomous experimentation.

8 Conclusion

A computational theory of discovery should explain which targets improving AI makes achievable and when research delivers them. The first suitable procedure need not give the earliest validated result. Additional target information improves some held-out forecasts and worsens others: the reasoning family, chemical endpoint and forecast budget change the inference. Aggregate success-rate agreement can also conceal large errors in particular chemical sources. These findings motivate conditional, testable connections between capability, information and validation. They support a research program for comparing scientific trajectories, not a validated calendar of breakthroughs.

AI use disclosure. A Codex AI assistant supported literature retrieval, argument development, mathematical exposition, code, checking and manuscript writing.

References

- Sanjeev Arora and Boaz Barak. Computational complexity: A modern approach, 2007. URL <https://theory.cs.princeton.edu/complexity/book.pdf>. Author web draft, January 8.
- Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. doi: 10.1073/pnas.1510507113.
- Jonathan Bragg, Mike D’Arcy, Nishant Balepur, et al. AstaBench: Rigorous benchmarking of AI agents with a scientific research suite. In *International Conference on Learning Representations*, 2026. URL <https://arxiv.org/abs/2510.21652v2>.
- Eric Chamoun, Yizhou Chi, Yulong Chen, Rui Cao, Zifeng Ding, Michalis Korakakis, and Andreas Vlachos. SciPaths: Forecasting pathways to scientific discovery, 2026. URL <https://arxiv.org/abs/2605.14600v1>. arXiv:2605.14600v1.
- Chaomei Chen, Yue Chen, Mark Horowitz, Haiyan Hou, Zeyuan Liu, and Donald Pellegrino. Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3(3): 191–209, 2009. doi: 10.1016/j.joi.2009.03.004.
- François Chollet. On the measure of intelligence, 2019. URL <https://arxiv.org/abs/1911.01547>. arXiv:1911.01547.
- C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.
- Payam Dibaeinia, Sudarshan Babu, Mei Knudson, Ali ElSheikh, Yibo Wen, Han Liu, Jason Perera, and Aly A. Khan. Position: Virtual cells need context, not just scale. In *International Conference on Machine Learning, Position Paper Track*, 2026. URL <https://icml.cc/virtual/2026/poster/67166>. Author version: bioRxiv 2026.02.04.703804.
- Florian E. Dorner, Yatong Chen, André F. Cruz, and Fanny Yang. Roc-n-reroll: How verifier imperfection affects test-time scaling. In *International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=3Gy5mmyuxn>.
- Mehmet Hamza Erol, Batu El, Mirac Suzgun, Mert Yuksekgonul, and James Zou. Cost-of-pass: An economic framework for evaluating language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=vC9S20zsgN>.
- Haosen Ge, Hamsa Bastani, and Osbert Bastani. Are AI capabilities increasing exponentially? a competing hypothesis, 2026. URL <https://arxiv.org/abs/2602.04836v2>. arXiv:2602.04836v2.
- Tilman Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.
- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y.
- Konstantin Hebenstreit, Constantin Convalexius, Stephan Reichl, Stefan Huber, Christoph Bock, and Matthias Samwald. What are the limits to biomedical research acceleration through general-purpose AI? *Scientific Reports*, 16:1304, 2026. doi: 10.1038/s41598-025-32583-w.
- Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer, 2021. URL <https://arxiv.org/abs/2102.01293v1>. arXiv:2102.01293v1.
- Michael Y. Hu, Jane Pan, Ayush Rajesh Jhaveri, Nicholas Lourie, and Kyunghyun Cho. Neural neural scaling laws, 2026. URL <https://arxiv.org/abs/2601.19831v1>. arXiv:2601.19831v1.
- Nathanael Jo and Ashia Wilson. Position: AI evaluations should be grounded on a theory of capability. In *International Conference on Machine Learning*, 2026. URL <https://arxiv.org/abs/2509.19590v2>. Author version: arXiv:2509.19590v2.

429 Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents
430 that matter. *Transactions on Machine Learning Research*, 2025. URL <https://openreview.net/forum?id=Zy4uFzMviZ>.
431

432 Kevin T. Kelly. *The Logic of Reliable Inquiry*. Oxford University Press, 1996. doi: 10.1093/oso/
433 9780195091953.001.0001.

434 Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, et al. Towards
435 a science of scaling agent systems, 2026. URL <https://arxiv.org/abs/2512.08296v3>.
436 arXiv:2512.08296v3; April 8.

437 Thomas Kwa, Ben West, Joel Becker, et al. Measuring AI ability to complete long software tasks.
438 In *Advances in Neural Information Processing Systems*, 2025. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2503.14499v4)
439 2503.14499v4. Updated version: arXiv:2503.14499v4.

440 Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. *Scientific Discovery:
441 Computational Explorations of the Creative Process*. MIT Press, 1987. URL [https://mitpress.](https://mitpress.mit.edu/9780262620529/scientific-discovery/)
442 [mit.edu/9780262620529/scientific-discovery/](https://mitpress.mit.edu/9780262620529/scientific-discovery/).

443 Fengyuan Liu, Jay Gala, Nilaksh, Dzmitry Bahdanau, Siva Reddy, and Hugo Larochelle. BRIDGE:
444 Predicting human task completion time from model performance, 2026a. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2602.07267v2)
445 [2602.07267v2](https://arxiv.org/abs/2602.07267v2). arXiv:2602.07267v2.

446 Yuxi Liu, Konpat Preechakul, Kananart Kuwarananchaoen, and Yutong Bai. The serial scaling
447 hypothesis. In *International Conference on Learning Representations*, 2026b. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2507.12549v4)
448 [2507.12549v4](https://arxiv.org/abs/2507.12549v4). Author version: arXiv:2507.12549v4.

449 Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I-Hung Hsu, Isaac Caswell,
450 Alex Pentland, Sercan Ö. Arık, Chen-Yu Lee, and Sayna Ebrahimi. ATLAS:
451 Adaptive transfer scaling laws for multilingual pretraining, finetuning, and decoding
452 the curse of multilinguality. In *International Conference on Learning Representa-*
453 *tions*, 2026. URL [https://proceedings.iclr.cc/paper_files/paper/2026/hash/](https://proceedings.iclr.cc/paper_files/paper/2026/hash/35c26a810471039b3427e524565cbef9-Abstract-Conference.html)
454 [35c26a810471039b3427e524565cbef9-Abstract-Conference.html](https://proceedings.iclr.cc/paper_files/paper/2026/hash/35c26a810471039b3427e524565cbef9-Abstract-Conference.html).

455 Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. Scaling laws are unreliable for downstream
456 tasks: A reality check. In *Findings of the Association for Computational Linguistics: EMNLP*
457 *2025*, pages 16167–16180, 2025. doi: 10.18653/v1/2025.findings-emnlp.877. URL [https://](https://aclanthology.org/2025.findings-emnlp.877/)
458 aclanthology.org/2025.findings-emnlp.877/.

459 Xiaoliang Luo, Hannan Saddiq, Rachel Asquith, Srishti Bhasin, Jamie McGraw, and Marc Warner.
460 A bayesian framework for forecasting and scenario analysis of ai task automation under scal-
461 ing constraints, 2026. URL [https://papers.ssrn.com/sol3/papers.cfm?abstract_id=](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7001782)
462 [7001782](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7001782). SSRN working paper.

463 Tyler H. McCormick. Position: Prioritize identifying structure, not complex models, for scientific
464 discovery. In *International Conference on Machine Learning, Position Paper Track*, 2026. URL
465 <https://arxiv.org/abs/2606.02632v1>. Author version: arXiv:2606.02632v1.

466 Elliot Meyerson and Xin Qiu. Position: Scaling LLM agents requires asymptotic analysis with
467 LLM primitives. In *Proceedings of the 42nd International Conference on Machine Learning*,
468 volume 267, pages 81797–81809. PMLR, 2025. URL [https://proceedings.mlr.press/](https://proceedings.mlr.press/v267/meyerson25a.html)
469 [v267/meyerson25a.html](https://proceedings.mlr.press/v267/meyerson25a.html).

470 Felipe Maia Polo, Seamus Somerstep, Leshem Choshen, Yuekai Sun, and Mikhail
471 Yurochkin. Sloth: scaling laws for llm skills to predict multi-benchmark perfor-
472 mance across families. In *Advances in Neural Information Processing Systems*, vol-
473 ume 38, 2025. URL [https://proceedings.neurips.cc/paper_files/paper/2025/](https://proceedings.neurips.cc/paper_files/paper/2025/hash/3834d037cd6e29ac7649df42611c5d66-Abstract-Conference.html)
474 [hash/3834d037cd6e29ac7649df42611c5d66-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2025/hash/3834d037cd6e29ac7649df42611c5d66-Abstract-Conference.html).

475 Bojana Ranković, Ryan-Rhys Griffiths, and Philippe Schwaller. Large language models as
476 uncertainty-calibrated optimizers for experimental discovery. *Nature Machine Intelligence*, 8:1466–
477 1477, 2026. doi: 10.1038/s42256-026-01283-z. URL [https://www.nature.com/articles/](https://www.nature.com/articles/s42256-026-01283-z)
478 [s42256-026-01283-z](https://www.nature.com/articles/s42256-026-01283-z).

479 Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the
480 predictability of language model performance. In *Advances in Neural Information Processing*
481 *Systems*, volume 37, 2024. URL <https://arxiv.org/abs/2405.10938v3>. Author version:
482 arXiv:2405.10938v3.

483 Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complex-
484 ity of optimistic exploration. In *Advances in Neural Information Processing Sys-*
485 *tems*, volume 26, 2013. URL [https://proceedings.neurips.cc/paper/2013/hash/](https://proceedings.neurips.cc/paper/2013/hash/41bfd20a38bb1b0bec75acf0845530a7-Abstract.html)
486 41bfd20a38bb1b0bec75acf0845530a7-Abstract.html.

487 Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learn-*
488 *ing: From Theory to Algorithms*. Cambridge University Press, 2014. URL
489 [https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/](https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf)
490 understanding-machine-learning-theory-algorithms.pdf.

491 Julian Shun. Parallel algorithms: Lecture 2, 6.506 algorithm engineering. Massachusetts Institute of
492 Technology, 2024. URL [https://jshun.csail.mit.edu/6506-s24/lectures/lecture2.](https://jshun.csail.mit.edu/6506-s24/lectures/lecture2.pdf)
493 pdf.

494 Joseph F. Traub, Grzegorz W. Wasilkowski, and Henryk Woźniakowski. *Information-Based Com-*
495 *plexity*. Academic Press, 1988.

496 Sang Truong, Yuheng Tu, Rylan Schaeffer, and Sanmi Koyejo. Item response scaling laws: A
497 measurement theory approach for efficient and generalizable neural scaling estimation. In *Interna-*
498 *tional Conference on Machine Learning*, 2026. URL <https://arxiv.org/abs/2606.07616v1>.
499 Author version: arXiv:2606.07616v1.

500 Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn.
501 Position: Will we run out of data? limits of LLM scaling based on human-generated data. In
502 *Proceedings of the 41st International Conference on Machine Learning*, volume 235. PMLR, 2024.
503 URL <https://proceedings.mlr.press/v235/villalobos24a.html>.

504 Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas,
505 Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright,
506 Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman
507 Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. Position: Evaluating
508 generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd*
509 *International Conference on Machine Learning*, volume 267, pages 82232–82251. PMLR, 2025.
510 URL <https://proceedings.mlr.press/v267/wallach25a.html>.

511 Arthur G. Werschulz. An overview of information-based complexity. Technical Report CUCS-022-02,
512 Columbia University, 2002. URL [https://mice.cs.columbia.edu/getTechreport.php?](https://mice.cs.columbia.edu/getTechreport.php?format=pdf&techreportID=152)
513 format=pdf&techreportID=152.

514 Parker Whitfill, Ben Snodin, and Joel Becker. Forecasting AI time horizon under compute slowdowns,
515 2025. URL <https://arxiv.org/abs/2511.19492v1>. arXiv:2511.19492v1.

516 Sean Wu, Pan Lu, Yupeng Chen, Jonathan Bragg, Yutaro Yamada, Peter Clark, David Clifton,
517 Philip Torr, James Zou, and Junchi Yu. Scientific reasoning does not reliably translate into
518 scientific forecasting in frontier AI, 2026. URL <https://arxiv.org/abs/2605.22681v2>.
519 arXiv:2605.22681v2.

520 Haotian Ye et al. Structured scaling of AI discovery across diverse scientific domains, 2026. URL
521 <https://arxiv.org/abs/2604.19341v2>. arXiv:2604.19341v2.

522 Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel machine:
523 Open-ended evolution of self-improving agents. In *International Conference on Learning Rep-*
524 *resentations*, 2026a. URL <https://arxiv.org/abs/2505.22954>. First released May 2025;
525 conference version March 2026.

526 Jenny Zhang, Bingchen Zhao, Wannan Yang, Jakob Foerster, Jeff Clune, Minqi Jiang, Sam Devlin,
527 and Tatiana Shavrina. HyperAgents, 2026b. URL <https://arxiv.org/abs/2603.19461v1>.
528 Version 1.

- 529 Pengsong Zhang, Heng Zhang, Huazhe Xu, Renjun Xu, Zhenting Wang, Cong Wang, Animesh Garg,
530 Zhibin Li, Arash Ajoudani, and Xinyu Liu. Scaling laws in scientific discovery with AI and robot
531 scientists, 2025. URL <https://arxiv.org/abs/2503.22444v2>. arXiv:2503.22444v2.
- 532 Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M. Collins, et al. Gen-
533 eral scales unlock AI evaluation with explanatory and predictive power. *Nature*, 652:58–
534 67, 2026. doi: 10.1038/s41586-026-10303-2. URL <https://www.nature.com/articles/s41586-026-10303-2>.
- 536 Yan Zhuang, Qi Liu, Zachary Pardos, Patrick C. Kyllonen, Jiyun Zu, Zhenya Huang, Shijin Wang,
537 and Enhong Chen. Position: AI evaluation should learn from how we test humans. In *Proceedings*
538 *of the 42nd International Conference on Machine Learning*, volume 267, pages 82483–82508.
539 PMLR, 2025. URL <https://proceedings.mlr.press/v267/zhuang25e.html>.

A Public-data prediction tests and reproduction

The empirical question is whether source capability measurements and an observable target descriptor predict outcomes beyond simpler accounts. The analyses reuse published aggregates; they do not repeat the original model evaluations. Table 3 maps the supporting evidence to the main claims.

Table 3: Supplement routing and evidence provenance.

Section	Scientific role and evidence type
A	Archive descriptor prediction, controls and uncertainty; inherited regression/PCA tools.
B	New exact recovery calculation using inherited binomial coverage.
C	Executed finite-domain example; observation restrictions determine search cost.
D	Premises and proofs for conditional capability and completion forecasts.
E	Campaign example, published-case audit and exact fixed-pilot analysis, including low power and evaluation costs.
F	Prior-work comparison and the proposed scientific connections.
G	Retrospective finite-table forecasting screen, including its negative result.

A.1 Data, preprocessing and scoring

The descriptor analysis uses the pinned public ObsScaling repository [Ruan et al., 2024]. Source measurements are aggregate rates on MMLU, ARC-C, HellaSwag, Winograd, TruthfulQA, GSM8K and XWinograd. The analysis joins complete source observations to the six BBH targets described below; missingness is not assumed random. The archive does not supply a common set of item-level outcomes or verified per-model release dates.

Within each training split, center and scale source metrics, extract three principal components (PCs) by singular-value decomposition, orient PC1 toward the standardized source mean, and scale PC coordinates to their training standard deviations. Fit this transformation on unique training models and apply it to the model–task rows. Test observations never determine the transformation. The source index is an observational predictor, not an exact reproduction of Ruan et al.’s full method. Published families can share ancestry, and all models share benchmark items; grouping by family cannot fully represent either dependence.

For observed aggregate rate y_i and prediction $\hat{p}_i$, rate-MSE is $N^{-1} \sum_i (y_i - \hat{p}_i)^2$ and fractional Bernoulli cross-entropy is $-N^{-1} \sum_i [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)]$. Predictions are clipped to $[10^{-8}, 1 - 10^{-8}]$ for scoring. Equal-cell scoring gives each model–task cell equal weight; the joint holdouts also report equal-family scores. Rate-MSE is not item-level Brier score: within-benchmark item variance and sampling uncertainty are absent. These aggregate losses do not certify calibration in a new population.

A.2 Descriptor design, rivals and interpretation

BBH logical deduction and shuffled-object tracking each have 3/5/7-object tasks. We select the archive’s three-shot chain-of-thought, strict-match rate columns and join model identifiers to the seven non-HumanEval source scores. The resulting panel has 73 models in 20 published families, each with six tasks: 438 observed cells. The groups are BLOOM, CodeLlama, DeepSeek-Coder, Falcon, GPT-Neo/J, Gemma, Llama, Llama-2, MPT, Mistral, Mixtral, OPT, Phi, Pythia, Qwen, Qwen1.5, StarCoder, StarCoder2, XGLM and Yi. Object count comes directly from task definitions. We did not label task difficulty from the outcomes or infer workflow depth from model performance.

For shared PC models, regress target logits, with rates clipped to $[\cdot005, \cdot995]$, on one or three PCs using ridge penalty 1 and an unpenalized intercept. The augmented model adds standardized log object count and its three interactions with PCs. Thus coefficients depend on a shared target descriptor without task-specific fitted intercepts. Each fit derives the PCA from unique training models, then applies it to the model–task rows. Count normalization also uses training rows. Primary tests hold

out one entire BBH task family at a time or one published model family at a time. An additional extrapolation test trains at 3/5 objects and tests all 7-object cells. Task-family holdout permits source measurements of those models; it tests target generalization, not new models simultaneously.

After the primary result, we added explicit exploratory rivals: count alone; PC1–3 with an additive count term; and $p = 1/n + (1 - 1/n)\sigma(\beta_0 + \beta^\top \mathbf{z})$, a chance-floor link. The last fits logits of clipped chance-adjusted rates and uses the same ridge penalty. It assumes the multiple-choice $1/n$ chance rate is a floor, although observed rates can fall below it. These controls distinguish a target-size effect from a need for capability–count interactions. The contract amendment identifies them as post-result tests, not independent confirmation. Coefficients and transformations use training data only; the ridge penalty is fixed.

Table 4: All descriptor-test scores, including secondary rivals. The target-family and model-family axes each contain 438 out-of-fold cells; the seven-object test contains 146.

Holdout	Predictor	Rate-MSE	Cross-entropy
larger_target	PC1	0.03074	0.48203
larger_target	PC3	0.03072	0.48263
larger_target	PC3_chance_floor	0.02172	0.46143
larger_target	PC3_count	0.01953	0.45884
larger_target	PC3_count_additive	0.01787	0.45529
larger_target	count_only	0.02786	0.48942
larger_target	pooled_mean	0.04458	0.52525
larger_target	task_family_mean	0.04466	0.52491
model_family	PC1	0.03201	0.56868
model_family	PC3	0.03301	0.57282
model_family	PC3_chance_floor	0.02703	0.55421
model_family	PC3_count	0.02376	0.54932
model_family	PC3_count_additive	0.02343	0.54858
model_family	count_only	0.03713	0.58280
model_family	pooled_mean	0.04275	0.59398
model_family	task_family_mean	0.04225	0.59272
target_family	PC1	0.03416	0.57559
target_family	PC3	0.03552	0.58015
target_family	PC3_chance_floor	0.02664	0.55396
target_family	PC3_count	0.02766	0.55988
target_family	PC3_count_additive	0.02600	0.55557
target_family	count_only	0.03947	0.58955
target_family	pooled_mean	0.04309	0.59487
target_family	task_family_mean	0.04309	0.59487

The additive count model has lower rate-MSE than the interaction model on all three separate holdout axes: rate-MSE .0260 versus .0277 on held-out task families, .0234 versus .0238 on held-out model families, and .0179 versus .0195 at seven objects. Count alone is worse than either capability-augmented count model. The chance-floor model also improves on capability-only prediction, and beats the interaction model on held-out task families. Thus these results support including simple task-size or chance information here, not the need for the richer interaction structure. The additive account has the lowest rate-MSE in these three comparisons; the joint holdouts below further challenge this choice. Paired differences resample the 20 published model families of fixed out-of-fold predictions, 2000 times. Against PC1–3, the primary count-interaction model’s rate-MSE differences are -0.00786 [-0.00963, -0.00608] for held-out task families; -0.00926 [-0.01180, -0.00691] for held-out model families; -0.01119 [-0.02642, 0.00523] for seven objects. These descriptive intervals omit refitting variability and uncertainty over new task families; overlapping training folds also induce dependence. With only two task families, we make no population-level significance or equivalence claim. Every model-level prediction, task-specific score, calibration-bin summary and negative comparison is retained.

A.3 Simultaneously unseen model and task families

Separate holdouts do not test whether the relation transports when both families are new. After the preceding results, we froze a follow-up challenge using the same archive and all eight predictors,

with no tuning. For each of the 20 model families and two task families, remove every row belonging to either held-out family from fitting; score only their intersection. These 40 folds cover each of the 438 cells once per predictor. A second set of 40 folds also restricts training to 3/5 objects and tests only seven-object intersections, covering 146 cells. PCA uses unique training models, and all transformations and regressions use training rows. Source measurements of test models remain required inputs.

Report both equal-cell losses and equal-family losses, the latter averaging each published model family’s mean loss with weight $1/20$. This checks whether unequal family sizes drive the conclusion; it does not make related training lineages independent. The code verifies disjoint training/test model and task families, unique test cells, excluded size, finite predictions and exact invariance to changing held-out target labels (640 predictor/fold checks).

Table 5: Follow-up joint holdouts. Lower is better; identical test cells and fixed predictors for each comparison. Equal-family scores average the 20 published model-family means. These are post-result archive challenges, not prospective or independent confirmation.

Predictor	Rate-MSE		Cross-entropy	
	Equal cell	Equal family	Equal cell	Equal family
Joint model/task holdout (438 cells)				
Pooled mean	0.04464	0.05323	0.59871	0.64146
Task-family mean	0.04464	0.05323	0.59871	0.64146
PC1	0.03570	0.04318	0.57954	0.62116
PC1–3	0.03790	0.04592	0.58748	0.63075
Count only	0.04102	0.05346	0.59388	0.65102
PC1–3 + count	0.02845	0.03558	0.56293	0.60588
PC1–3 + count interactions	0.03012	0.03771	0.56722	0.61125
PC1–3 + chance floor	0.02873	0.03503	0.55860	0.59759
Joint model/task holdout, seven objects (146 cells)				
Pooled mean	0.04857	0.04471	0.53373	0.56189
Task-family mean	0.04857	0.04471	0.53373	0.56189
PC1	0.03685	0.04091	0.49685	0.54539
PC1–3	0.03919	0.04406	0.50325	0.55314
Count only	0.02915	0.04253	0.49639	0.57769
PC1–3 + count	0.02088	0.02832	0.46907	0.53028
PC1–3 + count interactions	0.02385	0.03279	0.47760	0.54295
PC1–3 + chance floor	0.02899	0.03354	0.47951	0.53050

For joint family holdout, additive-count minus PC1–3 rate-MSE is -0.01425 for logical deduction and -0.00465 for shuffled-object tracking; equal-family differences are -0.01668 and -0.00399. For joint family and size holdout, additive-count minus PC1–3 rate-MSE is +0.00122 for logical deduction and -0.03784 for shuffled-object tracking; equal-family differences are +0.00755 and -0.03904. Negative differences favour the count descriptor. All per-task scores for the eight predictors and all five split axes are retained, including equal-family rescaling of the three earlier tests. These contrasts do not estimate uncertainty over a population of new task families.

Thus the descriptor gain under joint family holdout extends to both tasks, but the combined family-and-size test has a task-specific reversal. Moreover, equal-family rate-MSE favours the chance-floor predictor over additive count on joint family holdout, despite the opposite equal-cell ranking. The averaging population is part of the model-selection decision. Neither aggregate improvement licenses uniform task transport. The follow-up reuses previously inspected outcomes and cannot supply prospective or contamination-controlled confirmation.

A.4 Link sensitivity and full-pipeline refitting

These sensitivity analyses were specified after inspection of the earlier archive results. They reuse the same 73 models, 20 published families and two task families; they are neither new model evaluations nor independent confirmation. All five split definitions remain unchanged. Alongside PC1–3 and additive log count, test all quadratic capability terms, probit, rate-scale ridge regression clipped to $[0, 1]$, logit clipping at .001/.01, chance-floor plus count, and a free $1/n$ feature plus count. Ridge penalty remains 1, with an unpenalized intercept; transformations use training observations only. The

chance-floor variants fit clipped logits of $(y - 1/n)/(1 - 1/n)$ and transform back. Table 6 retains every predictor under the primary equal-cell loss. The artifact also contains equal-family scores and cross-entropy separately for both task families.

Table 6: Exploratory descriptor sensitivities: rate-MSE on identical cells. “Out” denotes exclusion from training; joint excludes both model and task families. No hyperparameter is selected on these holdouts.

Predictor	Task out	Model out	Size out	Joint	Joint + size
PC1–3, logit	0.0355	0.0330	0.0307	0.0379	0.0392
PC1–3 + count, logit	0.0260	0.0234	0.0179	0.0285	0.0209
Quadratic capability	0.0335	0.0396	0.0289	0.0439	0.0439
PC1–3, probit	0.0339	0.0320	0.0327	0.0362	0.0402
+ count, probit	0.0248	0.0227	0.0173	0.0272	0.0204
PC1–3, rate fit	0.0310	0.0304	0.0381	0.0333	0.0439
+ count, rate fit	0.0233	0.0223	0.0182	0.0256	0.0213
PC1–3, clip .001	0.0378	0.0344	0.0285	0.0404	0.0386
+ count, clip .001	0.0280	0.0246	0.0189	0.0308	0.0222
PC1–3, clip .01	0.0346	0.0325	0.0318	0.0369	0.0397
+ count, clip .01	0.0252	0.0230	0.0175	0.0275	0.0203
Chance floor	0.0266	0.0270	0.0217	0.0287	0.0290
Chance floor + count	0.0263	0.0261	0.0202	0.0284	0.0249
Chance feature + count	0.0260	0.0234	0.0179	0.0285	0.0209

Additive count improves on the matched capability-only fit across the tested links and clipping choices. Quadratic capability does not remove the pooled gain; chance-floor plus count does not consistently improve on the simpler additive predictor. The free chance feature scarcely changes the count result. Since n and $1/n$ are deterministically linked, none of these fits identifies whether reasoning burden or answer cardinality explains the gain. There are 211 below-floor cells out of 438: 78 in deduction and 133 in tracking. These aggregate observations diagnose a modeling restriction, not known individual-item failure probabilities.

For a separate uncertainty sensitivity, draw 20 model families with replacement in each of 500 replicates (seed 20260921). Keep all rows of each drawn family and exclude all copies of the test family when fitting a joint fold. Fit source standardization and PCA on unique models within each sampled family occurrence; refit count normalization and regression on the sampled training cells. Recompute out-of-fold predictions and paired additive-minus-PC1–3 losses for both joint designs. Equal-family summaries weight each sampled family occurrence equally; equal-cell summaries weight its cells. Table 7 reports the equal-cell percentiles; the artifact retains every draw and both weighting summaries. These percentiles describe model-family resampling of the entire fitted statistic, conditional on the two fixed task identities. They are not coverage guarantees for new target families, shared benchmark items, contamination or unobserved ancestry.

Table 7: Full-pipeline cluster-refit percentiles of additive-count minus capability-only rate-MSE; negative favors count. These differ from the earlier fixed-prediction intervals for count interactions.

Contrast	2.5th	Median	97.5th percentile
Joint / pooled	-0.0119	-0.0093	-0.0068
Joint / deduction	-0.0189	-0.0138	-0.0097
Joint / tracking	-0.0067	-0.0049	-0.0026
Joint + size / pooled	-0.0352	-0.0205	-0.0124
Joint + size / deduction	-0.0226	-0.0013	0.0113
Joint + size / tracking	-0.0589	-0.0401	-0.0250

The pooled gains persist, but the larger-task deduction interval spans zero under both weightings. Its originally observed worsening remains a descriptive result, not a stable population effect. Tracking drives the joint size gain. Repeated resampling cannot supply independent scientific-target evidence.

A.5 Design choices, artifacts and remaining validation

The logistic link makes an additive log-odds contrast interpretable and follows the cited response-model literature. It is not preferred by a universal theory; clipping/probit checks expose sensitivity near zero. Object count is reproducible and independent of our fitted outcomes, but it is a deliberately narrow descriptor. Clopper–Pearson intervals are chosen for finite-sample binomial coverage, with conservatism reported. Stipulated costs and yields span unchanged, worse and improved methods, with all grids retained rather than selected after inspecting outcomes.

The local analysis materials include the original protocol and subsequent amendments, data provenance, source-file hashes, scripts, results and manuscript sources. Running `python scripts/build_paper.py` regenerates analyses, tables, figures and PDF without network or model calls. The raw `ObsScaling` tables are retained with their Apache-2.0 license. Local availability is not a claim of public release. Archive holdouts do not remove training contamination, validate scientific outcomes or establish prospective prediction. Those require the matched target and temporal programme stated in the main text.

B Target-side identification and exact recovery

At source levels separated by known $\Delta > 0$, assume fixed probabilities $p_0, p_1 \in (0, 1)$ obey $\text{logit } p_1 - \text{logit } p_0 = \lambda \Delta$. Independent Bernoulli sampling within each level gives marginal binomial counts. Assigning .0125 error to each of four Clopper–Pearson tails yields simultaneous coverage at least .95 by a union bound. On that event, monotonicity of logit gives Equation 5. Cross-level dependence preserves this guarantee when both binomial marginals remain valid. Our recovery study additionally assumes independent levels to enumerate joint count probabilities.

For $N \in \{25, 100, 500\}$, $p_0 \in \{.01, .05, .2\}$, $\Delta \in \{.5, 1, 2\}$ and $\lambda \in \{0, .5, 1\}$, set $p_1 = \sigma(\text{logit } p_0 + \lambda \Delta)$. Sum coverage, positive-lower-bound probability and infinite-endpoint mass over all $(N + 1)^2$ counts with product-binomial weights. Record median interval width and cost $2Na$. All 81 settings are retained; an independent nested-loop implementation verifies coverage at all 27 settings with $N = 25$. The minimum grid coverage, 99.809%, is conservative relative to the analytical 95% lower bound. Zero-coupling settings are retained to check false exclusion.

Table 8: Exact recovery at coupling 1 and separation 2. Low initial success makes unbounded intervals more likely. The artifact retains all 81 settings, including zero coupling and smaller separations.

N	p_0	Δ	λ	Exclude zero (%)	Infinite (%)
25	0.01	2	1	0.000	81.456
25	0.05	2	1	4.905	27.759
25	0.20	2	1	44.744	0.380
100	0.01	2	1	6.585	36.651
100	0.05	2	1	90.659	0.592
100	0.20	2	1	99.967	0.000
500	0.01	2	1	97.210	0.657
500	0.05	2	1	100.000	0.000
500	0.20	2	1	100.000	0.000

At $p_0 = .05$, $\lambda = 1$, $\Delta = 2$, $p_1 = .2800$. Median interval widths at $N = 25, 100, 500$ are 4.26, 1.69 and .71; measurement costs are $50a, 200a, 1000a$. Greater separation improves information only while a common affine-logit relation and fixed protocol remain credible. No finite sample guarantees a finite interval for every possible outcome. More than two levels and held-out target observations are needed to challenge model form; between-system differences cannot isolate capability from training data or other system changes.

Why two exact rates do not establish a finite horizon. Let source capability follow $z(t) = t \geq 0$. Compare target log-odds $g_1(z) = -3 + z$ and $g_2(z) = -3 + 2z/(1 + z)$. At $z = 0, 1$, both give success probabilities $\sigma(-3) = .0474$ and $\sigma(-2) = .1192$: even exact observations cannot distinguish them. Both responses increase, but $g_2(z) = -1 - 2/(1 + z) < -1$, so its success probability stays below $\sigma(-1) = .2689$. At $q = .8$, the first crosses at $3 + \log 4 = 4.39$ scenario-years and the second never crosses. At a withheld third level $z = 2$, they instead predict .2689 and .1589. This

illustrative construction shows why two-level slope estimation requires the affine-logit assumption for extrapolation. A third observation can challenge it, but does not establish the response at every future level.

C Finite-domain synthesis and information access

The target is an executable program matching $f_a(x) = ax \bmod 2^n$ for every $x \in \{0, \dots, 2^n - 1\}$, with odd hidden a uniform over its 2^{n-1} values. The permitted candidates use the same program grammar. The rejection-only interface hides the target outputs and returns one bit after a complete domain comparison; it supplies no counterexample input. Distinct coefficients differ at $x = 1$, so exactly one candidate is valid. Success here requires a positive verifier response; returning an untested candidate does not count. Before success, every rejected candidate only excludes itself.

Write $M = 2^{n-1}$. Conditional on any negative-query history, the hidden coefficient remains uniform over the unqueried candidates. Every fresh-query policy therefore encounters it at a uniformly distributed position in $1, \dots, M$; randomized policies are mixtures with the same distribution. Success by k distinct checks has probability k/M , giving the exact $\lceil qM \rceil$ cap and mean $(M + 1)/2$. This is the standard one-positive search calculation [Russo and Van Roy, 2013, cf. Example 2]. If an untested final guess were accepted, success could instead reach $\min\{(k + 1)/M, 1\}$; that changes the declared success criterion.

The implementation enumerates all odd coefficients and all inputs at bit widths 2 through 8, checks that only the hidden coefficient passes, and counts all input evaluations without early exit. At $n = 2, 4, 6, 8$, candidate counts are 2, 8, 32 and 128; .9-reliability caps are 2, 8, 29 and 116, giving work 8, 128, 1856 and 29,696. For a uniform random candidate order, mean calls are $(2^{n-1} + 1)/2$. With a target-output query at input 1, the coefficient is observed directly; one query plus a full check costs $1 + 2^n$ evaluations. These count observations and evaluations as equal units for illustration; wall-clock costs can differ.

A concrete requirement profile is $(n, 2^{n-1}, \text{observation interface}, 2^n)$. Changing the interface removes the search bottleneck without changing arithmetic capability. This is an executed, restricted program-synthesis example, not a lower bound against procedures with richer specification access or an empirical LLM scaling result. Full finite-domain verification establishes exactly the specification chosen here, not an unbounded input guarantee.

D Formal conventions and conditional forecasts

Feasible sets and conditional forecasts separate procedure randomness, uncertainty about a response model, and uncertainty about technological development. This section supplies the formal premises used in the main paper.

Procedures and information. A procedure includes model weights, inference budget, tools, shared memory, agent coordination, stopping, human assistance and the observation protocol. The information state K records both initial evidence and permitted access to further observations; it is not merely a count of training examples. Equation 1 permits adaptive acquisition within the resource allowance. Observational data and interventional data need not be interchangeable. Feasibility describes existence in the declared procedure class, not a guarantee that an evaluator finds the best procedure. Under matched resources and access, the crossing of any admissible procedure is therefore an upper bound on the class-level capability horizon. Equality follows for a singleton class; it otherwise requires establishing the attainable envelope. The worked affine-logit scenarios use singleton classes.

When unused resources are optional, $\mathcal{B}_Q$ is upward closed. Every procedure feasible under $\mathbf{b}$ remains admissible under a componentwise larger allowance. This says nothing about an individual agent’s response to a larger prompt or an enforced increase in reasoning. Enlarging the procedure class by freely optional methods likewise cannot shrink the attainable set. Costs of developing or installing the new method must be included when the comparison concerns the whole research programme.

Capability horizon and actual completion. Equation 3 is a first-passage definition along an entire system-development path. It need not be monotone if access, targets or available procedures change.

747 The infimum need not be attained for an arbitrary path; in the worked continuous monotone examples
 748 it is attained. Reliability is evaluated within a campaign resource vector that includes its allowed
 749 duration. A threshold crossing concerns when a suitable campaign becomes possible, not when a
 750 campaign actually produces a result.

751 A completion forecast additionally specifies launch times, stopping, resource allocation and validation
 752 duration. A system below a reliability threshold can succeed, and one above it can fail on an
 753 individual run. The conditional campaign probabilities below connect these distinct quantities
 754 without introducing a second capability measure.

755 **From campaign success to the first result.** Condition on evidence D and deployed policy π .
 756 For a fixed finite ordered schedule of completion dates, let A_j be failure of campaign j . On
 757 each reached all-failure history define $1 - h_{Q,j} = \Pr(A_j \mid A_1, \dots, A_{j-1}, D, \pi)$. The chain rule
 758 gives $\Pr(A_1 \cap \dots \cap A_n \mid D, \pi) = \prod_{j=1}^n (1 - h_{Q,j})$, proving Equation 7 for n completions by T .
 759 Independence is a sufficient reason for conditional and marginal campaign probabilities to coincide,
 760 not a premise of the identity. If an earlier all-failure event has probability zero, discovery has
 761 already occurred almost surely under the model and subsequent factors are immaterial. Uncertain
 762 completion schedules require averaging the full event probability over the schedule and development
 763 path. Likewise, conditional-on-parameter predictions must be averaged jointly; multiplying separately
 764 averaged success probabilities generally changes the forecast.

765 D.1 Conditional horizon calculations and verification

766 Let $h = \text{logit } q - \eta_0 > 0$ be the remaining log-odds gap. For positive v, λ , solving $\eta_0 + \lambda vt = \text{logit } q$
 767 gives $\tau = h/(\lambda v)$. If $\lambda = 0$ and initial success is below q , no crossing occurs. The same is true for
 768 $\lambda < 0$ with positive v under this particular constant-coefficient scenario. These are mathematical
 769 limits of the assumed paths, not evidence that a target can never be solved after a method or
 770 information change.

771 For a rectangular parameter region $h \in [h_-, h_+]$, $\lambda \in [\lambda_-, \lambda_+]$ and $v \in [v_-, v_+]$, with every lower
 772 endpoint positive, monotonicity gives

$$\frac{h_-}{\lambda_+ v_+} \leq \tau \leq \frac{h_+}{\lambda_- v_-}. \quad (8)$$

773 The rectangle specifies which parameter combinations are allowed; it does not assume probabilistic
 774 independence. If constraints couple the parameters, optimize the same expression over that region.
 775 If the region includes $\lambda = 0$, it admits infinite horizon. A hard access gate opening at time t_E
 776 makes success zero before t_E ; for a monotonically increasing post-access curve, the horizon is
 777 $\max\{t_E, h/(\lambda v)\}$. The ungated and access-gated examples therefore do not have identical initial
 778 target success, as stated in Figure 2.

779 The constructed nonidentification example uses three ungated models with $\lambda = 1, .5, 0$, identical
 780 source path, and identical target success at zero. If the evidence contains only source observations
 781 and that initial target probability, each agrees exactly with it. Their future target probabilities and
 782 crossings disagree. Positive source–target evidence at additional capability levels, appropriate priors
 783 or a structural restriction could distinguish them. Neither a theorem about universal unpredictability
 784 nor a claim about how often these worlds occur follows.

785 **Proof of the time-error bound.** Let I contain τ and $\hat{\tau}$, with $\ell(\tau) = \hat{\ell}(\hat{\tau}) = \text{logit } q$. Suppose ℓ is
 786 differentiable on I and $\ell'(t) \geq m > 0$ throughout, and $|\ell(t) - \hat{\ell}(t)| \leq \varepsilon$ there. Then

$$m|\hat{\tau} - \tau| \leq |\ell(\hat{\tau}) - \ell(\tau)| = |\ell(\hat{\tau}) - \hat{\ell}(\hat{\tau})| \leq \varepsilon.$$

787 This standard inverse-sensitivity argument proves Equation 6. It assumes both crossings exist in the
 788 interval. It does not follow from pointwise confidence intervals or calibration averaged over observed
 789 models. A plateau at the threshold, an unmodeled jump or a crossing outside I can invalidate its
 790 premises. Uniform control can instead give a set of possible crossing times by inverting lower and
 791 upper performance bands, with simultaneous rather than pointwise coverage.

792 **Numerical reproduction.** `scripts/analyze_capability_horizons.py` computes the four
 793 paths, analytic and numerically solved crossings, the positive-coupling interval and time-error

examples. It also checks that the ungated models have equal initial performance, that the source trajectory is identical, and that the two-world identification example cannot exceed worst-case 1/2 accuracy. The plots show assumed functions, not fits. These scenario checks are distinct from the public-data fits in Appendix A.

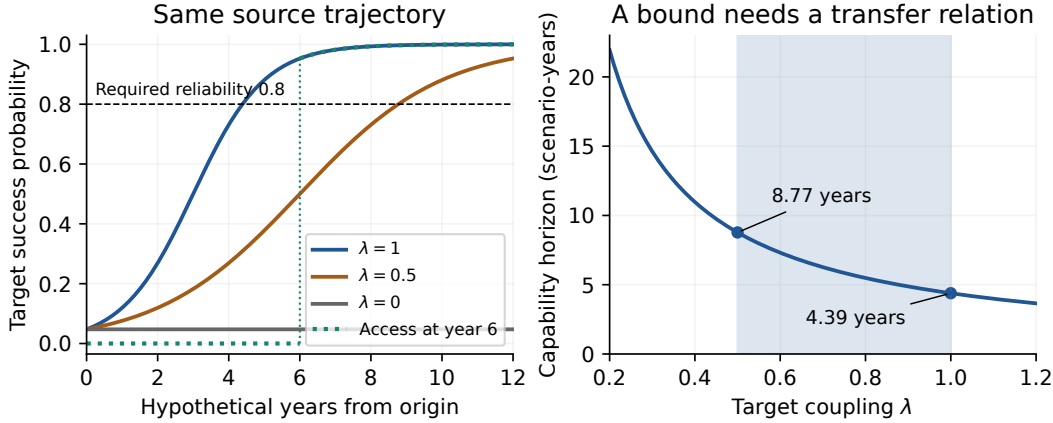


Figure 2: Constructed capability horizons, not scientific forecasts. All four scenarios share the source path $x(t) - x_0 = t$. Left: target success under strong coupling ($\lambda = 1$), weaker coupling ($.5$), no coupling (0), or strong coupling with required access unavailable until year 6. The access-gated scenario starts at zero; the other three start at $.0474$. Right: horizon sensitivity to positive λ at $q = .8$, $v = 1$ and $\eta_0 = -3$. The shaded range corresponds to the stipulated bound $\lambda \in [.5, 1]$, not a statistical confidence interval. A calendar interpretation requires evidence for these parameters and the growth path.

Capability and completion reversal. For independent serial attempts of duration d and success probability $p > 0$ launched at s , the number of attempts to first success is geometric, with mean $1/p$. Expected completion is therefore $s + d/p$. In the main assay example, launching A and B at their respective availability dates gives means $\tau_A + 4/.8 \approx 9.39$ and $\tau_B + .25/.8 \approx 9.09$ scenario-years. Subtracting the two means shows that B finishes earlier on average exactly when $d_A - d_B > q(\tau_B - \tau_A)$, because both deployed versions have fixed success q . This compares two separate deployment choices. Adding B as an optional procedure does not force an optimal policy to abandon A; the policy can retain whichever choice is preferable. The validation durations and response curves are stipulated, not measured biomedical forecasts.

E Campaign reliability and the cost of evidence

The following analyses connect target success probabilities to a concrete adoption decision. They show how a target-response probability enters experimental resource requirements and how uncertain that response may remain. Their stipulated parameters cannot be substituted for the source–target coefficients in the main horizon example.

E.1 When is more reasoning worth its cost?

Consider a deliberately simplified candidate-selection procedure. Each attempt costs c units to propose a candidate and v units to test it. Both are hypothetical monetary costs, not time; shared setup costs are set to zero. Let $p(c)$ be the probability that a proposed candidate meets the declared binding and cell-assay criteria after validation. The notation allows reasoning expenditure to affect yield but does not assume that it improves it. Hold the target and testing protocol fixed, and assume independent candidate outcomes with this same probability, where $0 < p(c) < 1$.

To find at least one validated hit with probability at least ρ , where $0 < \rho < 1$, reserve enough budget for n attempts. Failure on all attempts has probability $(1 - p(c))^n$. Thus the smallest attempt cap and

its budget are

$$n_\rho(c) = \left\lceil \frac{\log(1 - \rho)}{\log(1 - p(c))} \right\rceil, \quad B_\rho(c) = n_\rho(c)(c + v). \quad (9)$$

This standard repeated-trial calculation gives the reserved budget for this procedure; stopping after a hit can spend less. It does not estimate elapsed time or optimize over other methods.

Table 9: Hypothetical choices for reaching at least 90% probability of one validated hit. Assay cost $v = 9$ in every row; outcomes are independent with constant yield. Shared activity, batch effects or adaptive selection can violate this model. Costs share arbitrary monetary units; setup is excluded here. Yields are stipulated, not estimates from cited studies.

Procedure	Proposal cost c	Hit probability p	Attempts n	Budget B
Baseline	1	0.10	22	220
Cheaper proposals	0.1	0.10	22	200.2
Better selection	2	0.20	11	121
Costlier, unchanged yield	2	0.10	22	242

Table 9 shows the decision. Tenfold cheaper proposals save 9% of the campaign budget. Doubling proposal expenditure while doubling validated yield saves 45%, because fewer assays are needed. Doubling expenditure without improving yield instead raises the budget by 10%. Neither cheaper computation nor a higher source benchmark score determines which row applies. That requires evidence about $p(c)$ on the biomedical target.

The threshold that changes the decision. At a new per-attempt cost of 11, a strict saving over the baseline budget 220 permits at most 19 attempts. Reaching $\rho = .9$ therefore requires $p \geq 1 - .1^{1/19} \approx .1141$. The yield need not double: increasing it from .10 to at least .1141 suffices under this model. A pilot should resolve whether the new procedure clears that threshold, rather than merely whether its source score improves. Uncertainty matters: a point estimate above the threshold is not a reliability guarantee. Appendix E.3 propagates a yield interval into budget bounds.

Independence is not essential for a sufficient guarantee. If each candidate succeeds with conditional probability at least $p_{\min}$ after every reached failure history, the chain rule bounds n consecutive failures by $(1 - p_{\min})^n$. The same attempt cap is then sufficient, though not necessarily minimal. An average marginal yield is weaker: if all candidates share one binary outcome, repeated tests never raise success above p . An expert-filtered selector has a yield conditional on that whole selection procedure, not on raw proposals. Adaptive selection can also change yield after each observation. Neither selected-hit counts nor average marginal yield alone supply the history-conditional guarantee.

E.2 A published case: what is measured, and what is missing?

Co-Scientist reports both computational scaling and laboratory outcomes [Gottweis et al., 2026]. Table 10 records what connects its hypothesis-ranking result to its initial acute myeloid leukemia (AML) drug-repurposing experiment. The six fields distinguish the source measurement, target, adapter, baseline, outcomes and costs. They document the narrower workflow comparison; they are not a replacement for capability-to-target calibration.

The combined system-and-expert workflow produced active candidates. The authors distinguish this early validation from clinical efficacy. What remains unmeasured here is how changing the source capability changes yield and cost under a fixed selection-and-testing protocol. The table states the scope of this audit.

Why the missing comparison matters. Even knowing a new procedure’s yield need not settle adoption. Stipulate its yield is .20 and each attempt costs 11: its .9-reliability budget is 121. If baseline yield is .10 at cost 10, the baseline budget is 220 and adoption saves money. If baseline yield is .30 at the same cost, its budget is 70 and adoption costs more. Both worlds agree on the new procedure’s results. An unmeasured baseline therefore leaves even the direction of savings unresolved. These are constructed worlds, not estimates for Co-Scientist.

Table 10: Connecting capability evidence to a scientific endpoint: a bounded audit of Co-Scientist Fig. 2a and initial AML Fig. 3a–c. Unknown entries identify measurements needed for the forecast, not evidence of zero benefit.

Field	Reported evidence and supported boundary
Source measurement	Top-hypothesis Elo ratings rise across ten temporal buckets covering 203 research goals (Fig. 2a); ratings are relative comparisons, not assay probabilities.
Biomedical target	Initial AML repurposing: cell-viability inhibition (Fig. 3a–c), distinct from the hypothetical binding-plus-cell-activity target in this appendix and from clinical benefit.
Selection and people	Oncologists selected five drugs from thirty ranked candidates before testing; expert filtering is part of the procedure.
Matched baseline	These results do not pair each reasoning budget with a matched laboratory arm and baseline procedure.
Outcome denominator	Three of five selected drugs showed reported inhibition. This does not estimate unfiltered generator yield or how yield changes with reasoning budget.
Costs and decision	Complete matched campaign costs and reuse savings are not identified by these results. Unknown benefit is not zero benefit.

The pilot below uses complete outcome denominators and explicit uncertainty under its stipulated observation model. A discovery-stage claim needs validation of its declared endpoint, not a clinical trial unless patient benefit is claimed.

E.3 Propagating uncertainty into campaign reserves

How uncertainty changes the budget. For a fixed independent-attempt procedure, $B_\rho(p)$ decreases as p increases. Thus a valid interval $0 < p_L \leq p \leq p_U < 1$ implies

$$(c + v) \left\lceil \frac{\log(1 - \rho)}{\log(1 - p_U)} \right\rceil \leq B_\rho(p) \leq (c + v) \left\lceil \frac{\log(1 - \rho)}{\log(1 - p_L)} \right\rceil. \quad (10)$$

For measured p , use a lower confidence bound appropriate to the prespecified sampling and selection procedure. On the event that it covers the true fixed p , substituting that bound provides the stated conditional campaign guarantee. Confidence coverage and campaign reliability are separate probabilities and should be reported separately; a .9 reliability target is not automatically a .9 unconditional guarantee after estimation. A zero lower bound gives no finite cap through this argument. An ordinary binomial interval for a marginal yield cannot establish the uniform history-conditional bound needed by an adaptive policy. If both procedures' yields or costs are uncertain, propagate both; robust savings require the new upper budget bound below the baseline lower bound.

E.4 Exact decision analysis for a fixed pilot

This analysis asks how much evidence the proposed comparison can require, within the same hypothetical campaign model. It does not estimate biological yields or validate the reporting framework against real teams. The analysis contract was saved before outcomes: two frozen procedures, baseline and candidate per-attempt costs 10 and 11, campaign reliability $\rho = .9$, pilot sizes $N \in \{25, 50, 100, 250, 500, 1000\}$ per arm, baseline yield .10, and candidate yields .08, .10, .12, .20, .30. The range includes worse, unchanged, near-threshold and substantially improved procedures; all results are retained.

Pilot and future campaign are different stages. A pilot always completes its prespecified N independent attempts in each arm; it does not stop at the first hit or choose its size from interim results. Each arm has a fixed unknown yield p_j and known cost a_j , with $j = 0$ for baseline and $j = 1$ for candidate. For this calculation the arms are independent too. All selected attempts enter the denominator; validity and failure rules, target, candidate-selection procedure and assay are fixed in advance. The subsequent campaign seeks at least one hit and reserves $B_j(p) = a_j \lceil \log(1 - \rho) / \log(1 - p) \rceil$, with $B_j(0) = \infty$ and $B_j(1) = a_j$. Setup costs are zero here. Shared batches, adaptive candidate selection, changing yield or uncertain costs need another observation/cost model.

888 **A conservative decision rule from standard intervals.** Let $X_j \sim \text{Binomial}(N, p_j)$. Give each
889 arm an equal-tailed Clopper–Pearson interval $[L_j, U_j]$ with coverage at least $1 - \alpha/2$, where $\alpha = .05$
890 [Clopper and Pearson, 1934]. For count $0 < x < N$, the endpoints are beta quantiles

$$L(x) = F_{\text{Beta}(x, N-x+1)}^{-1}(\alpha/4), \quad U(x) = F_{\text{Beta}(x+1, N-x)}^{-1}(1 - \alpha/4).$$

891 Use $L(0) = 0$, $U(N) = 1$, and the same quantile formulas for the other endpoint at these extremes.
892 With probability at least $1 - \alpha$, both true yields are covered: apply the union bound to the two
893 noncoverage events. Monotonicity then gives $B_j(U_j) \leq B_j(p_j) \leq B_j(L_j)$ simultaneously. Report

adopt: $B_1(L_1) < B_0(U_0)$,
non-saving: $B_1(U_1) \geq B_0(L_0)$,
unresolved: otherwise.

894 On the joint coverage event either resolved conclusion is correct. Thus the unconditional probability
895 of a wrong resolved conclusion is at most α for every pair of fixed yields under this model. This is not
896 a 95% probability of correctness conditional on selecting a conclusion, nor the future campaign’s 90%
897 reliability. The guarantee is inherited confidence-interval reasoning, not a novel statistical method.
898 The rule can be conservative; no optimality or comparison against all decision rules is claimed.

899 **Exact enumeration and comparator.** For each prespecified parameter setting, sum the decision
900 indicator over all $(N + 1)^2$ possible count pairs weighted by their independent binomial probabilities.
901 The implementation evaluates equivalent binomial tails, with direct double-sum cross-checks at
902 $N = 25$. There are no random seeds, Monte Carlo error bars or fitted parameters. The point-
903 estimate comparator adopts when $B_1(X_1/N) < B_0(X_0/N)$ and otherwise keeps the baseline. It
904 uses identical observations and cost accounting, but has no stated error guarantee and never defers.
905 Report adoption, non-saving, unresolved and wrong-decision probabilities, rather than comparing
906 a selective rule only on its accepted cases. A separate numerical check over $N \in \{25, 100, 500\}$,
907 $p_0 \in \{.05, .10, .20\}$ and 41 equally spaced $p_1 \in [.02, .40]$ verifies the error inequality at 369 settings;
908 the analytical coverage argument, not that finite grid, establishes the guarantee.

909 **Sensitivity to reliability and sampling error.** Reliability ρ is the desired chance of a future
910 campaign hit; α bounds a wrong resolved decision over repeated pilots under the sampling model.
911 They answer different questions. A retrospective exact sensitivity crosses $\rho \in \{.8, .9, .95\}$, $\alpha \in$
912 $\{.01, .05, .10\}$ and candidate yields $p_1 \in \{.10, .12, .20, .30\}$, with baseline yield .10, $N = 100$
913 per arm and costs 10/11 fixed: 36 settings. All count pairs are weighted by independent binomial
914 probabilities. At $\rho = .9$ and $p_1 = .20$, adoption is .646%, 4.572% and 10.902% as α increases; for
915 unchanged yield, false adoption is .0005%, .0182% and .0714%. At $\alpha = .05$ and doubled yield,
916 adoption is 5.099%, 4.572% and 5.003% across the three ρ values. Integer campaign caps make
917 these differences nonmonotone. Low power persists in this grid, but the grid neither justifies a user’s
918 threshold nor compares all decision rules. Thresholds should reflect the consequences of error and
919 deferral; prespecification alone supplies no such justification. The standard-library implementation in
920 `scripts/analyze_pilot_threshold_sensitivity.py` inverts binomial tails, checks coverage
921 and reproduces the original four overlapping settings to absolute error below 10^{-10} . All 36 outcomes
922 are retained; this is hypothetical analysis, not empirical confirmation.

923 **Evaluation costs and boundaries.** Pilot expenditure is $N(10 + 11) = 21N$, charged separately
924 from the future campaign. For the stipulated $.10 \rightarrow .20$ improvement, future budgets fall from 220
925 to 121. If the entire pilot is charged as an additional expense and its own hits receive no credit,
926 offsetting that expense in reserved-budget savings requires $\lceil 21N/99 \rceil$ otherwise identical future
927 campaigns. This assumes the saving is known and realized in subsequent budget allocations; it is
928 not an expected cash-payback calculation. Campaigns may stop early, so actual expenditure differs
929 from the reserve. This is an accounting illustration, not an estimate of experimental prices or net
930 scientific value. Pilot hits may themselves be valuable; deployment frequencies and task changes also
931 matter. The comparison shows why an uncertainty-aware evaluation can be sound yet uneconomic
932 for a one-off decision. Biological dependence or distribution change can invalidate the calculation;
933 no amount of computation validates its premises.

934 `scripts/analyze_decision_pilot.py` saves every setting, checks interval endpoints against
935 SciPy’s exact binomial interval API, verifies campaign caps directly from success probabilities, and
936 generates Figure 3. The complete build reruns these checks. This numerical analysis supports the

937 implications of the stipulated model; a prospective matched study is still needed before claiming
938 practical advantage in biomedical research.

Table 11: All 30 prespecified hypothetical pilot settings. Baseline yield is .10 and its true .9-reliability campaign budget is 220; B_1 is the candidate’s true budget. Costs are arbitrary monetary units. Adoption probabilities refer to repeated pilots under the stipulated yields. The interval rule may defer; the point-estimate rule does not. At $p_1 \leq .10$, adoption is erroneous. Percentages are rounded to three decimals; displayed zeros can be small positive probabilities. Pilot expenditure is separate from campaign reserves.

N	p_1	B_1	Interval adopt (%)	Unresolved (%)	Point adopt (%)	Pilot cost
25	0.08	308	0.000	99.994	30.720	525
25	0.10	242	0.001	99.996	40.456	525
25	0.12	209	0.003	99.995	49.851	525
25	0.20	121	0.157	99.843	78.978	525
25	0.30	77	2.412	97.588	95.106	525
50	0.08	308	0.001	99.934	29.846	1050
50	0.10	242	0.006	99.969	43.245	1050
50	0.12	209	0.023	99.967	56.165	1050
50	0.20	121	0.911	99.089	89.335	1050
50	0.30	77	12.630	87.370	99.094	1050
100	0.08	308	0.003	99.630	25.069	2100
100	0.10	242	0.018	99.900	41.877	2100
100	0.12	209	0.088	99.895	58.853	2100
100	0.20	121	4.572	95.428	95.801	2100
100	0.30	77	46.899	53.101	99.952	2100
250	0.08	308	0.001	98.704	13.432	5250
250	0.10	242	0.013	99.836	36.103	5250
250	0.12	209	0.170	99.817	63.892	5250
250	0.20	121	28.697	71.303	99.687	5250
250	0.30	77	96.783	3.217	100.000	5250
500	0.08	308	0.000	95.678	5.694	10500
500	0.10	242	0.008	99.694	29.528	10500
500	0.12	209	0.299	99.692	68.445	10500
500	0.20	121	71.254	28.746	99.994	10500
500	0.30	77	99.996	0.004	100.000	10500
1000	0.08	308	0.000	83.945	1.242	21000
1000	0.10	242	0.004	99.281	22.175	21000
1000	0.12	209	0.630	99.365	75.618	21000
1000	0.20	121	98.503	1.497	100.000	21000
1000	0.30	77	100.000	0.000	100.000	21000

939 F Foundations and further research directions

940 Computational discovery builds on models of law and concept discovery [Langley et al., 1987],
941 reliable inquiry [Kelly, 1996], and transformative discoveries in knowledge networks [Chen et al.,
942 2009]. Information-based complexity combines costly observations with computation [Traub et al.,
943 1988, Werschulz, 2002]. We build on these foundations to study the evolving reach of AI research
944 procedures.

945 Measurement validity is part of this question. Wallach et al. [2025] distinguish conceptualization
946 from operational measurement; Zhuang et al. [2025] advocates psychometric evaluation, and Chollet
947 [2019] emphasizes priors and generalization difficulty. A descriptor needs a defensible meaning as
948 well as predictive accuracy. Cost-controlled agents and reproducible research environments supply
949 relevant evaluation practice [Kapoor et al., 2025, Bragg et al., 2026].

950 Scientific scaling is already an agenda [Zhang et al., 2025, Ye et al., 2026], alongside resource-
951 constrained automation and scientific-pathway forecasting [Luo et al., 2026, Chamoun et al., 2026].
952 Data-stock forecasts illustrate uncertainty under resource scenarios [Villalobos et al., 2024]; scientific

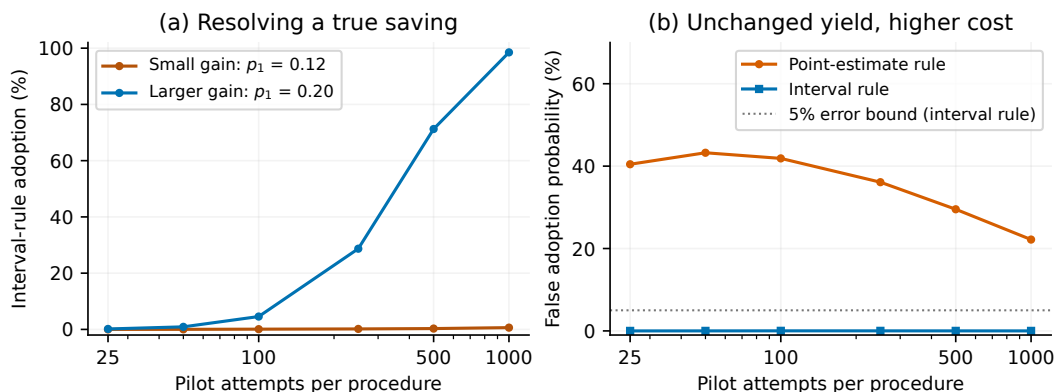


Figure 3: Exact hypothetical pilot analysis. Baseline yield is .10, per-attempt costs are 10 and 11, and future campaign reliability is .9. Left: probability of recommending adoption when a true saving exists. The interval rule requires the candidate upper budget bound below the baseline lower bound. Right: probability of false adoption when yield is unchanged and costs rise; the two panels answer different questions. Lines join evaluated sizes, not a fitted scaling law. Neither rule is claimed optimal.

Table 12: Closest predictive foundations and the remaining scientific question. The right column states obligations for our programme, not established advantages.

Foundation	Established construction	Additional question here
Observational scaling [Ruan et al., 2024]	Shared capability coordinates predict downstream performance.	Which target requirements improve prediction beyond these coordinates?
Demand and response scales [Zhou et al., 2026, Truong et al., 2026]	Model ability and item demands jointly determine responses.	Which descriptors transport to unseen research targets?
Task-duration horizons [Kwa et al., 2025]; compute scenarios [Whitfill et al., 2025]	Estimate reliable task duration and condition its growth on resources.	What extra relation connects that ability to a specified scientific endpoint?
Information complexity and workflow analysis [Werschulz, 2002, Meyerson and Qiu, 2025]	Account for observation access, computation and dependencies.	How do changed procedures alter capability and completion times separately?

953 event forecasting differs empirically from scientific reasoning [Wu et al., 2026]. Our question is
 954 which measurements justify moving from capability to target performance, then to a date. Neither
 955 logistic fitting nor computational discovery is new.

956 Learned representations can change experimental search [Ranković et al., 2026], and discovery loops
 957 can use evaluated histories [Ye et al., 2026]. Recursive self-improvement may alter capability growth
 958 or procedure-development costs [Zhang et al., 2026a,b]. A successful cycle supports that cycle;
 959 sustained acceleration requires measurements across successive improvements and their full costs.

960 G An exploratory finite experimental-table forecast screen

961 We test the proposed next step on public tables distributed with GOLLuM [Ranković et al., 2026]:
 962 four additive-screening plates, five Buchwald–Hartwig reaction tables, one Suzuki–Miyaura table
 963 and one C2-yield table. These are eleven tables from four source families, not eleven independent
 964 scientific domains. The repository snapshot is pinned in the artifact. Its Apache license is retained;
 965 underlying experimental provenance remains attributed to the source study. We do not rerun the
 966 laboratory experiments or reproduce GOLLuM’s language-model optimizer.

967 The original replay target was an observed objective at or above the admitted table’s 90th percentile.
 968 All predictors received that threshold. After dropping incomplete and duplicate input rows, we

uniformly subsampled at most 512 candidates without consulting outcomes. All eleven admitted sets contain 512 candidates. Numeric inputs are scaled by catalogue ranges and strings treated as categorical identities; outcome, procedure text and derived-output columns are excluded. The percentile uses the admitted archive to define a benchmark target, not information proved available to a historical forecaster. Ties and source dependence are reported in the retained metadata.

We ran 50 seeded campaigns per table. Five initially uniform observations were followed by a fixed Gaussian-process upper-confidence policy, with acquisition mean plus one posterior standard deviation, noise .01 and kernel $\exp(-2d)$, where d averages squared normalized numeric differences and categorical mismatches. Standardization of outcomes uses only observations already revealed. Every measured candidate costs one unit; observations within the pilot count against the total budget. No language-model training or paid inference is performed.

At 5 and 10 observations, we predicted at least one hit by budget 20 (primary) or 40 (sensitivity). We compared ridge-logistic forecasts using budget alone, adding catalogue size, adding the best observed gap to threshold, and adding pilot mean, dispersion and pairwise kernel similarity. Catalogue size cannot differentiate these equally sized sets.

For pilot outcomes y_i and supplied threshold u , let $s = \max\{|u|, \text{sd}(y_i), 10^{-8}\}$, using population standard deviation. Best gap, mean gap and dispersion are $(\max_i y_i - u)/s$, $(\bar{y} - u)/s$ and $\text{sd}(y_i)/s$. Similarity averages the off-diagonal kernel entries among pilot candidates. These calculations use revealed outcomes and the known catalogue.

All predictors assign probability one after an observed pilot hit; logistic fits use only training campaigns without such a hit. The 550 campaigns yield 2,200 prefix–budget records (two prefixes and two budgets), evaluated by four predictors; these are repeated views of the same campaigns, not 8,800 independent trials. At pilot five, 221 campaigns have already hit and 329 remain. All other features use only the pilot and declared threshold; held-out future outcomes are evaluation labels. Each evaluation withheld an entire source family, keeping all five Buchwald–Hartwig tables or all four additive plates together. Forecast features were standardized on training families only. Losses were averaged equally within tables and then across source families, with the not-yet-hit stratum retained separately.

Table 13: Held-out-family Brier scores after five pilot observations, predicting success by budget 20. Lower is better. All forecasts use the same 550 campaign traces; pilot discoveries count as successes.

Source family	Budget-only	+ best gap	+ pilot summaries
additives	0.00107	0.00175	0.00307
buchwald–hartwig	0.01186	0.01213	0.01208
c2–yield	0.01931	0.01971	0.01980
suzuki–miyaura	0.03868	0.03946	0.03950

Across families, budget-only Brier score is .01773, adding the best gap gives .01826, and pilot summaries give .01862. Thus the additional pilot descriptors do not improve the primary score. On campaigns without an initial hit, scores are .03106, .03197 and .03253. By budget 20 the family-averaged hit rate is 98.2%; all retained campaigns have a hit by budget 40. This ceiling leaves little uncertainty for a discoverability forecast to resolve. The result neither validates the proposed descriptor nor establishes equivalence or impossibility; it identifies an uninformative easy-target regime for this screen. The stricter-endpoint analysis below is exploratory: its endpoints were chosen after this primary result was known.

For reference, uniform sampling without replacement from a table with N candidates and H archive hits has hit probability $1 - \binom{N-H}{b} / \binom{N}{b}$ by budget b . The artifact retains this oracle-informed reference using full-table hit counts; it is not a forecast of the adaptive policy.

The artifact includes all candidate IDs, observed paths, source hashes, forecast features, fold predictions, per-table losses and calibration bins. Unique-query and future-outcome permutation checks protect the observed-prefix protocol. Four heterogeneous source families do not justify a calibrated target-population interval. This retrospective table replay supplies neither an intervention test of AI capability nor an information-available-at-the-time discovery forecast. A subsequent study would need a new independently fixed endpoint with substantial unresolved outcomes and held-out target families, not a redefinition selected to improve these scores.

G.1 Exploratory comparisons with simple forecasts

Near-universal success can make a low Brier score easy to obtain. We therefore added three reference forecasts after seeing the original replay: always predict success; predict success only if the pilot already found a hit; and use success prevalence from training families. These reuse every campaign and the original endpoint. They are a retrospective challenge, not independent confirmation.

For the prevalence forecast, withhold the whole test source family. At each pilot/budget setting, use only other families' campaigns without a pilot hit; average their binary outcomes within tables, then tables within families and families equally. Assign this rate to unresolved test campaigns and probability one to those already known to have hit. Every fit retains three training families; no empty training stratum occurs. The always-success forecast uses no fitted parameters. The observed-hit-only forecast represents no further progress after the pilot, not a proposed competitive method.

Table 14 compares Brier score with log-loss, which penalizes confident errors more heavily. Both use the original scoring convention: clip probabilities to $[10^{-6}, 1 - 10^{-6}]$, average campaigns within tables, then tables within families, then families equally. The all-campaign column includes 550 campaigns; the no-pilot-hit column includes 329. Training prevalence nearly matches the budget-only logistic forecast. Always-success also nearly matches its Brier score but has substantially worse log-loss: rare failures contradict its near-certainty. The two metrics therefore expose different weaknesses. Neither result validates a prospective discovery forecast.

Table 14: Post-result baseline challenge at pilot five and budget 20. Lower is better for both losses. All rows use identical campaign outcomes and family/table weighting; the three added references are marked *. The two strata overlap and are not replications.

Forecast	All campaigns		No pilot hit	
	Brier	Log-loss	Brier	Log-loss
Budget-only logistic	0.01773	0.08549	0.03106	0.14865
+ pilot summaries	0.01862	0.10181	0.03253	0.17661
Training prevalence*	0.01774	0.08571	0.03109	0.14900
Always success*	0.01800	0.24868	0.03160	0.43654
Observed hit only*	0.57025	7.87830	0.96840	13.37897

The comparison covers all four pilot/budget settings, both strata and each source family. Numerical checks reproduced the original scores; perturbing held-out outcomes left fitted predictions unchanged, confirming their exclusion from fitting. These checks use the same 550 campaigns and provide no additional evidence of generalization to new sources.

G.2 Endpoint sensitivity and source-level contrasts

Following the 98.2% hit rate at the 90th-percentile endpoint, we chose stricter 95th- and 99th-percentile thresholds for an exploratory analysis. Their forecast scores were computed afterward. All 550 saved forty-step paths are reused without rerunning or changing the selection policy. Each endpoint redefines pilot-hit status, future-hit labels and the threshold-relative features. We refitted the same ridge-logistic budget and summary predictors with training-only standardization and leave-one-source-group exclusion. The existing best-gap predictor was then added as a simpler challenge to the 99th-percentile gain. All thresholds remain full-table hindsight information. This ordering makes the follow-up exploratory; it is not a new confirmation set.

Tables 15 and 16 retain all endpoint/pilot/budget comparisons and the primary source differences. Within each fold, the training-only prevalence reference balances tables and source groups among no-hit pilots and assigns one to already-hit pilots. Scores average campaigns within tables, tables within source groups, and the four groups equally, as in the original experiment. The same campaign appears in several rows; rows are not independent experiments.

At the 99th percentile, pooled log-loss falls from .76525 to .49995; the best-gap and prevalence references score .88142 and .76174. At the 95th percentile it worsens from .31881 to .38814. Omitting any one source from the average of the existing held-out losses gives 99th-percentile Brier differences from $-.14301$ to $-.10520$. This is reweighting of fixed fold results, not refitting with only two training sources. No target-population confidence interval is claimed. The no-pilot-hit 99th-percentile

Table 15: Endpoint sensitivity: family-balanced Brier scores. All three targets, two pilots, two budgets, four predictors and both strata are retained. The two strata overlap. Lower is better.

Percentile	Pilot/budget	Budget	Best gap	Summaries	Prevalence
All campaigns					
90	5/20	0.01773	0.01826	0.01862	0.01774
90	5/40	0.00000	0.00000	0.00000	0.00000
90	10/20	0.01730	0.01768	0.01786	0.01734
90	10/40	0.00000	0.00000	0.00000	0.00000
95	5/20	0.09346	0.10043	0.10460	0.09418
95	5/40	0.00503	0.00514	0.00515	0.00504
95	10/20	0.08097	0.08762	0.09705	0.08119
95	10/40	0.00509	0.00508	0.00519	0.00507
99	5/20	0.28716	0.34035	0.16489	0.28510
99	5/40	0.22316	0.22688	0.29591	0.22966
99	10/20	0.21259	0.24759	0.14558	0.21475
99	10/40	0.23235	0.23634	0.27048	0.22517
No pilot hit					
90	5/20	0.03106	0.03197	0.03253	0.03109
90	5/40	0.00000	0.00000	0.00001	0.00000
90	10/20	0.09101	0.09376	0.09401	0.09089
90	10/40	0.00002	0.00001	0.00002	0.00000
95	5/20	0.11963	0.12798	0.13341	0.12071
95	5/40	0.00599	0.00613	0.00614	0.00600
95	10/20	0.21364	0.23029	0.25620	0.21459
95	10/40	0.01277	0.01274	0.01303	0.01271
99	5/20	0.30416	0.36056	0.17473	0.30196
99	5/40	0.23720	0.24108	0.31476	0.24408
99	10/20	0.26740	0.30743	0.18903	0.26974
99	10/40	0.26631	0.27181	0.29886	0.25758

Table 16: Paired Brier difference, summaries minus budget-only, at pilot five/budget 20. Negative favors summaries. These are four descriptive source contrasts, not independent draws from a specified scientific population.

Source	90th	95th	99th
Additive screens	+0.00201	+0.02455	-0.17351
Buchwald–Hartwig	+0.00022	+0.00714	-0.06007
C2-yield	+0.00050	+0.00172	-0.11672
Suzuki–Miyaura	+0.00083	+0.01113	-0.13881

score also improves, from .30416 to .17473. Gains cannot be attributed solely to the deterministic already-hit gate. They do not persist at the forty-observation budget: at pilot five, summaries score .29591 versus .22316 for budget-only; at pilot ten, .27048 versus .23235. Thus the follow-up identifies an endpoint-and-budget-specific contrast, not uniform summary superiority.

An independent solver reproduced the original predictions, and held-out-label perturbations left fitted predictions unchanged; source-file identities and thresholds were also checked. The saved 26,400 forecast records reuse the same 550 campaign histories and include every table/group loss and log-loss summary. The endpoint-dependent reversal warrants prospective testing of what early observations predict; it does not establish an AI-capability effect, clinical utility or a fitted discovery date.

G.3 Budget-specific forecasts and opposing source errors

The longer-budget reversal admits a simple rival explanation: perhaps one set of summary coefficients cannot describe both budgets. We tested this after the endpoint analysis, retaining all three endpoints, both pilots, both budgets and the same four source exclusions. No trajectories or outcomes were added. The shared model is the original logistic regression. Separate models fit each budget independently, using both pilot lengths. For each held-out source, feature centering and scaling are still computed on

the complete outer training set, so changing budgets does not also change feature units. Budget is constant within a separate fit; its penalized coefficient is zero. The best-gap predictor receives the same comparison.

The primary separate fits retain penalty 1. Each budget has half as many training rows, so a second control uses penalty $n_h/n_{\text{pool}} = 1/2$: with weights summing to the training-row count, this preserves the penalty per weighted observation. Intercepts remain unpenalized. This changes coefficient sharing and parameter count; it is neither a new forecasting algorithm nor an isolated causal intervention on information. The same deterministic probability-one rule applies after a pilot hit. Predicted probabilities are clipped to $[10^{-6}, 1 - 10^{-6}]$ for both losses, as in the original analysis.

Table 17: Exploratory forecast-specification check. Brier scores use all campaigns and equal source/table weighting. Shared: original coefficients across budgets. Separate: independent budget fits, penalty 1. Matched: separate fits with penalty 1/2. All endpoints, pilots and budgets are shown; these are repeated forecasts of the same 550 histories.

Endpoint	Pilot	Budget	Budget-only			Summaries		
			Shared	Separate	Matched	Shared	Separate	Matched
90	5	20	0.01773	0.01774	0.01774	0.01862	0.01872	0.01874
90	5	40	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
90	10	20	0.01730	0.01735	0.01737	0.01786	0.01791	0.01784
90	10	40	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
95	5	20	0.09346	0.09340	0.09341	0.10460	0.10526	0.10533
95	5	40	0.00503	0.00504	0.00504	0.00515	0.00607	0.00671
95	10	20	0.08097	0.08096	0.08095	0.09705	0.09944	0.09954
95	10	40	0.00509	0.00507	0.00508	0.00519	0.00513	0.00522
99	5	20	0.28716	0.28494	0.28495	0.16489	0.15424	0.15423
99	5	40	0.22316	0.22607	0.22606	0.29591	0.33922	0.34428
99	10	20	0.21259	0.21396	0.21395	0.14558	0.13438	0.13436
99	10	40	0.23235	0.22957	0.22958	0.27048	0.31398	0.32052

At the 99th-percentile endpoint, pilot five and budget 20, separate summary fits improve Brier score from .16489 to .15424. At budget 40 they worsen it from .29591 to .33922; log-loss worsens from .99753 to 2.31803. The matched-penalty score is .34428 (log-loss 2.50883). The separate best-gap reference also worsens at budget 40 (.24185 versus the shared .22688). Separate fitting therefore does not remove the observed long-budget failure, but this check does not exclude other representations or calibration procedures.

True hits are nested: a history that hits by observation 20 also hits by 40. Across all three endpoints and both pilots, none of the three original predictors violates this ordering in its forecasts. Separate summary fits do so for 2/550 histories at the 95th percentile and 10/550 at the 99th, both at pilot five. Matching the penalty increases these counts to 5 and 16; the matched best-gap fit has one strict-endpoint violation. All other settings have zero. These alternative fits are diagnostic comparisons, not recommended coherent completion distributions.

The apparent agreement of the overall success-rate mean has a different limitation. For the original strict-endpoint summary forecast at pilot five/budget 40, the equally weighted source means are predictions .34964, .99113, .96927 and .96732, against observed rates 1, .976, .28 and .98 for additives, Buchwald–Hartwig, C2 and Suzuki–Miyaura. Underprediction and overprediction cancel: their averages are .81934 and .809. For each campaign, the residual is predicted hit probability minus its binary outcome. Within each table, mean squared residual equals squared mean residual plus residual variance. Averaging this identity by the scoring weights gives $.29591 = .22732 + .06860$; rounding accounts for the last digit. Thus 76.8% of this descriptive error comes from within-table mean residuals. This is not a calibration guarantee or a population bias estimate; it diagnoses the existing predictions without fitting to held-out labels. Averaging sources before squaring would hide these opposing errors.

Checks reproduced all 19,800 original logistic predictions exactly and verified eleven source hashes, 180 held-out-label perturbations, nested outcomes and the error decomposition. Retained forecasts and losses cover both score strata and every source/table. The 59,400 records reuse 550 histories, not new experiments; independently chosen endpoints and sources remain necessary before generalizing.