The Natural Order Hypothesis for LLMs

Levi Neuwirth*

13 November 2026

Abstract

We propose a natural order hypothesis: language models trained on substantially shared information become able to solve problems reliably in approximately the same order, with training methods chiefly changing how quickly and how far learning proceeds. We postulate this weak form and conjecture a strong one, in which a single capability scale would also govern how often different labs' models make the same discovery. Evidence from open mathematical conjectures and chess training runs shows that models' solved sets are nested, as the weak form requires, and supports a common order only in part; the strong form has little support. We give one dated forecast and propose tests that separate the forms.

1 Introduction

OpenAI and Anthropic have published separate papers reporting refutations of the Exponential Time Hypothesis (OpenAI, 2026e; Claude and Anthropic, 2026). Both accounts date a discovery milestone to October 16, 2026: a natural-language disproof in one and a core algorithm in the other (OpenAI, 2026a; Claude and Anthropic, 2026). The matching dates are self-reports of different events, and we do not treat them as evidence. They raise a question that can be studied without them: why might models from different labs come to solve the same problems, and in what order?

We propose a *natural order hypothesis*: language models trained on substantially shared information, including its representation, tend to learn to solve problems reliably in approximately the same order. Training methods chiefly change the pace and extent of acquisition. The name borrows from second-language acquisition, where a similar distinction concerns the sequence of learning and its response to instruction (Krashen, 1982).

We state the hypothesis in two forms (Section 2). The weak form is the claim about order, and it is what we postulate. The strong form adds a single capability scale on which every model sits, so that comparable models succeed on the same problems at similar rates. We state it as a conjecture and expect it to fail when read exactly. The strong form implies the weak, and the weak implies that the sets of problems different models can solve are nested. The three claims call for different observations: nesting can be seen in a single cross-section, order only by following acquisition through training, and a common scale only by comparing solve probabilities.

Shared learning orders already have precedents in neural-network research (Hacohen et al., 2020); Liu et al. (2026b) explicitly propose a predictable compositional order during language-model pretraining. The question here extends to open mathematical problems and the possible connection between acquisition order and discovery across labs.

The evidence reaches the three claims unevenly. On a bank of open conjectures, solved sets are nearly nested across labs and across successive releases. Order among newly acquired problems is

*Work carried out while a student at the Technical University of Denmark (DTU).

supported on one bank of chess positions followed through training, at a rank correlation near .76 among the pairs observed inside both windows. It is contradicted on a second bank across input formats, and it is unobserved for language models, where release ladders are coarse and each verdict is a single attempt. The strong form has little support. A direct test requires repeated observations of abilities acquired during training, alongside the problems already solved or still beyond reach.

The two forms answer different parts of the opening question. The weak form says which problems come next for a lineage at a given stage. Only the strong form says how often a second lab's model succeeds where a first has, and neither says that two labs attempt a problem in the same week. We give one dated forecast of nested reach and propose tests that separate the forms (Section 8).

2 The hypothesis

A model may solve a problem occasionally before it can do so reliably. We propose that the order in which reliable solutions become possible transfers across model lineages.

We state the proposal in two forms. The weak form concerns that order alone and is the one we postulate. The strong form adds a single scale of capability and is offered as a conjecture. Unqualified, *natural order hypothesis* means the weak form.

Natural order hypothesis (weak form)**.** Gradient-trained language-model lineages exposed to substantially shared informational inputs, including how those inputs are represented, acquire the ability to solve problems in an approximately common order. Within a specified domain and under fixed evaluation conditions, problems that become reliably solvable earlier in one lineage tend also to become reliably solvable earlier in others. Differences in training methodology chiefly affect the pace and extent of acquisition, with comparatively limited changes to its order. The hypothesis permits differences in success probability, acquisition speed, and the intervals between successive acquisitions.

Training methodology means how training is conducted given the informational input and its representation. A test must define the eligible differences between learners before observing their order. Which architecture, size and optimization differences to include remains a design question (Section 8.1). Commercial releases with opaque training inputs are observational comparisons; we call successive releases from one provider a *release ladder*, without assuming a documented training genealogy.

The weak form predicts that acquisition order transfers across eligible lineages. Similar aggregate performance alone does not require this: two models can solve equally many problems while acquiring different ones first.

2.1 What the terms commit to

Solve probability refers to repetitions of a fixed procedure: the model, problem statement, hints, agent scaffold, tools, compute budget, and solution-acceptance rule.

Let $p_{\ell i}(k)$ be the probability that lineage ℓ, at checkpoint k, solves problem i under the fixed evaluation conditions. For a fixed observation endpoint T_ℓ, define *stable acquisition* at threshold $q = .5$ by

$$\tau_{\ell i}(q; T_\ell) = \min\{k : p_{\ell i}(j) \geq q \text{ for every observed } j = k, \ldots, T_\ell\}.$$

This is the first observed checkpoint from which solve probability remains at or above the threshold through the observation endpoint. A problem reliably solvable throughout the window is already

acquired when observations begin (*left-censored*). If its solve probability is below the threshold at the final checkpoint, no stable acquisition is observed within the window (*right-censored*), even if it was reliably solvable earlier. Two problems in the same boundary category cannot be ordered from these observations alone. Acquisition is relative to the chosen endpoint: extending the window can change it after a later loss.

To compare two lineages, we ask which problem in each pair was acquired first. Within a lineage, a pair is *ordered* when the observations place one acquisition before the other by more than a stated margin. Problems acquired closer together than the margin are left unordered, as are two problems in the same boundary category. The margin is the resolution of the measurement and is fixed before orders are compared. In our release-ladder data it is one rung, so that conjectures first solved at the same release are tied; in our chess data it is the interval-separation rule of Appendix C. A pair is *resolved* when it is ordered in both lineages. If one acquires X before Y and the other Y before X, this is a *reversal*. The reversal rate is the fraction of resolved pairs whose orders disagree. *Coverage* is the fraction of all problem pairs under study that are resolved. For example, if 40 of 100 pairs are resolved and four of those disagree, coverage is 40% and the reversal rate is 10%. This compares sequences without requiring the lineages to share a training clock.

A reversal rate r corresponds to a rank correlation of $\gamma = 1-2r$ over the resolved pairs. Unrelated orders give $r = .5$ and $\gamma = 0$; $r = .1$ gives $\gamma = .8$.

Approximately has a meaning only against two references. One is the reversal rate between unrelated orders, .5. The other is the rate between two runs of a single training recipe that differ only in random seed, which no comparison of methods can be expected to beat. The weak form asserts that reversal rates between eligible lineages, among pairs resolved at the stated margin and with adequate coverage, lie near the second reference and far from the first. Neither of our data sets supplies the seed reference; the direct test in Section 8.1 requires it.

Agreement must also be measured among acquisitions observed inside both windows, since boundary comparisons can agree without any new acquisition. We call a resolved pair *interior* when both of its problems are acquired inside the observation window in both lineages, and a *boundary* pair otherwise. Single-attempt records provide proxies for acquisition; estimating the underlying probabilities requires repeated attempts and uncertainty treatment (Section 8.1).

2.2 What the weak form implies for solved sets

Write $A_\ell(k) = \{i : \tau_{\ell i} \le k\}$ for the set of problems that lineage ℓ has acquired by checkpoint k. At the endpoint, $A_\ell(T_\ell)$ is the set of problems reliably solvable there.

Proposition 1 (Nested reach)**.** *For two lineages ℓ and ℓ' and any checkpoints k and k', the number of pairs reversed at margin zero is at least*

$$|A_\ell(k) \setminus A_{\ell'}(k')| \cdot |A_{\ell'}(k') \setminus A_\ell(k)|.$$

In particular, if the two lineages have no reversals, one of $A_\ell(k)$ and $A_{\ell'}(k')$ contains the other.

Proof. Take i in $A_\ell(k)$ but not in $A_{\ell'}(k')$, and j in $A_{\ell'}(k')$ but not in $A_\ell(k)$. In ℓ, problem i is acquired by checkpoint k and problem j is not, so i precedes j; this holds whether i is left-censored, j is right-censored, or neither. In ℓ', by the same argument, j precedes i. Every such pair is therefore a reversal, and there are as many of them as the product. Without reversals, one of the two differences is empty. □

With a positive margin the same argument applies to problems separated from the cut by more than the margin.

Nested reach is therefore what the weak form looks like in a single cross-section, or along a release ladder with few rungs. It is necessary for the weak form and falls well short of it. Nested sets order only the pairs that some observed cut separates. Two problems that both lineages acquire between the same consecutive observations are left unordered by nesting, and their order is what the weak form adds. Single-attempt solved sets are proxies for acquired sets, so observed nesting supports this consequence only to the extent that single verdicts track reliable solving.

2.3 The strong form: a shared capability scale

The strong form makes two additions. On the item side, problems keep one order by solve probability in every lineage and at every checkpoint, and not only at the moment each crosses a threshold. On the model side, models are themselves ordered: a single coordinate places every checkpoint of every lineage, and two models at the same position succeed on each problem at the same rate. We expect this form to fail when read exactly. We state it because it is the form that would connect a common order to how often discoveries repeat.

Natural order hypothesis (strong form)**.** Problem-solving performance in gradient-trained language models exposed to substantially shared informational inputs, including how those inputs are represented, is approximately governed by a common capability scale. Within a specified domain and under fixed evaluation conditions, a model's position on that scale and a problem's accessibility jointly determine the probability that the model solves the problem through a common function increasing in both arguments. Differences between lineages and training methodologies chiefly affect models' positions on the scale. Models at comparable positions consequently exhibit approximately similar solve probabilities on individual problems.

An exact version makes the relationship explicit:

$$p_{\ell i}(k) = F\big(c_\ell(k), a_i\big).$$

Here $c_\ell(k)$ is the lineage's capability at checkpoint k, and a_i is the accessibility of problem i. The same function F applies to every lineage and problem. Increasing capability raises solve probability for a given problem; increasing accessibility raises it for a model at a given capability. The item side is the monotonicity of F in accessibility. The model side is that capability is a single number and F is shared, so that models at the same position on the scale have the same solve probability on each problem.

The item side imposes a common ordering of problems by solve probability. If $a_i \geq a_j$, then in every lineage and at every checkpoint,

$$p_{\ell i}(k) \geq p_{\ell j}(k).$$

Problem i remains at least as likely to be solved as problem j throughout training. In psychometrics, *invariant item ordering* means that test items retain the same ordering by probability of a correct response across ability levels: item response functions do not cross as latent ability varies (Sijtsma and Junker, 1996, p. 83). The exact strong form imposes this property across model lineages through the shared function F.

Proposition 2 (The strong form implies the weak)**.** *Suppose the item side of the exact strong form holds: there are accessibilities a_i such that $a_i \geq a_j$ implies $p_{\ell i}(k) \geq p_{\ell j}(k)$ in every lineage and at every checkpoint. Then no pair of problems is reversed between any two lineages, at any threshold q, or between two different thresholds.*

Proof. Fix a lineage, an endpoint and a threshold, and let $a_i \geq a_j$. Suppose problem j stays at or above the threshold from some checkpoint through the observation endpoint. Problem i, whose solve probability is at least as high at every checkpoint, meets the threshold throughout the same period, so its stable acquisition occurs no later: $\tau_{\ell i} \leq \tau_{\ell j}$. Now suppose i strictly precedes j in some lineage at some threshold. Then $a_j \geq a_i$ is impossible, so $a_i > a_j$, and in every lineage and at every threshold i is acquired no later than j. □

The argument uses only the item side, and it holds even if capability fluctuates during training, since it compares the two probabilities checkpoint by checkpoint. Together with Proposition 1 it gives three claims, each implied by the next: nested reach, the weak form, and the strong form. Evidence against an earlier claim is evidence against every later one.

The converse of Proposition 2 fails. The weak form allows two lineages to acquire X before Y and later become more reliable on Y; the exact strong form excludes that change in probability order. The weak form is stated at one threshold, and the order of crossing at .5 need not be the order at .25 or .75. And the weak form does not ask that models with equal overall performance have equal probabilities on each problem.

Proposition 2 relates the exact forms. An approximately common scale does not by itself bound the reversal rate: where many problems have nearly equal accessibility, small departures in probability reorder their threshold crossings. This is why the weak form is stated with a margin, so that pairs the measurement cannot separate are unresolved and are counted neither as agreements nor as reversals. Testing an approximate strong form still requires a measure and tolerance for departures from the common scale, both of which remain open. A particular parametric model can also impose more than the strong form. The Rasch model specifies an additive logistic F, so rejecting it does not reject the strong form, and failing to reject it does not establish a common scale (Section 5.1).

2.4 What separates the forms

Table 1 sets out which observations bear on which claim. Three of them are compatible with the weak form and count against the strong: an acquisition order that changes with the threshold, a probability order that reverses after both problems are acquired, and models matched on overall performance that differ on individual problems.

The forms also differ in what they say about discoveries made twice. The weak form says which problems come next. Two lineages at nearby stages face the same frontier, and if both attempt it they attempt problems of similar standing. It does not say that two lineages are at nearby stages, since it permits any difference in pace, and it does not say how often a second attempt succeeds. Aligned acquisition at .5 gives two independent single attempts a joint success probability of .25.

The model side of the strong form supplies the missing quantity. For conditionally independent attempts by matched models with common item-specific probability p, averaging over a specified population of problems gives

$$\Pr(Y_B = 1 \mid Y_A = 1) = \frac{\mathbb{E}[p^2]}{\mathbb{E}[p]} = \mathbb{E}[p] + \frac{\mathrm{Var}(p)}{\mathbb{E}[p]}.$$

Replication is frequent when accessibility is polarized, with most problems either well within reach or well beyond it, and rare when success is uniformly unlikely: if every eligible item has $p = .01$, the rate is .01. The strong form supplies no universal replication rate. A numerical forecast needs a specified population and its own empirical justification.

Table 1: Observations and the claims they bear on. Each claim is implied by the one to its right, so an observation against a claim is against every claim to its right.

Observation	Nested reach	Weak form	Strong form
Solved sets nest; interior order unobserved	Supports	Necessary only	Necessary only
Reliably solved sets cross	Against	Against	Against
Interior pairs reverse well beyond the seed reference	Compatible	Against	Against
Interior order agrees to near the seed reference	Supports	Supports	Necessary only
Acquisition order changes with the threshold	Compatible	Compatible	Against
Probability order reverses after both problems are acquired	Compatible	Compatible	Against
Models matched on total score differ on individual problems, while reliably solved sets still nest	Compatible	Compatible	Against
Matched models have matched probabilities on each problem	Silent	Silent	Supports

A double on the same day therefore asks for more than either form. The lineages must be at nearby stages, both must attempt the problem at nearby times, and the problem must lie where both succeed often.

2.5 An order and its possible explanations

One possible explanation of the weak form is that accessibility reflects a problem's relation to training information: shared inputs would produce similar orders, while differences in relevant information would move particular problems. On this account the hypothesis says more than that some problems are harder than others. The order should survive a change of training method, and it should move when the information or its representation changes. An account based on problem structure alone predicts the first and not the second.

The shortest sufficient hint could measure a problem's relation to training information, provided missing knowledge can be distinguished from elicitation, since short opening cues can substantially change mathematical performance (Zhang et al., 2026; Wang et al., 2026b). The proposed ETH hint study would examine assistance for one theorem and route. Other explanations include problem structure, common training practices, contamination and convergent representations (Huh et al., 2024); Section 7 considers how to distinguish them.

3 Background

3.1 Orders already observed in neural learning

A common order in neural learning has direct precedents. Hacohen et al. (2020) report similar orders of example learning across neural-network architectures in the settings they study, while also constructing cases in which agreement fails. Choshen et al. (2022) investigate grammatical development across language models, and Chen et al. (2023b) discuss a natural order of skills in data selection. Shared learning order is therefore an existing empirical question. The present proposal concerns its extension to reliable problem solving, particularly open problems, and the evidence needed to connect an acquisition order to multiple discovery.

The Implicit Curriculum Hypothesis of Liu et al. (2026b) is an especially close predecessor: it concerns the order in which constructed tasks emerge during language-model pretraining. Their cross-model agreement is appreciably higher under fixed performance thresholds than relative thresholds. The corpora share sources, including in the Amber and Crystal models released through LLM360 (Liu et al., 2023). Both the overlap in inputs and the sensitivity to the acquisition criterion bear directly on the present tests.

Structured learning dynamics (Saxe et al., 2019; Nakkiran et al., 2019), word acquisition (Chang and Bergen, 2022), and mathematical skill development (Mishra et al., 2025) offer further precedents. Representational alignment (Li et al., 2016; Huh et al., 2024) and correlated errors (Goel et al., 2025; Kim et al., 2025) concern related forms of agreement at an endpoint.

3.2 A natural order in second-language acquisition

Krashen's natural order hypothesis proposes a predictable sequence of grammatical acquisition (Krashen, 1982), while work on processability and instruction distinguishes the pace of learning from its route (Bettoni and Di Biase, 2015). Evidence also points to the learner's background: Murakami and Alexopoulou (2016) find stronger agreement within than between seven first-language groups, and earlier case evidence complicates accounts based on frequency (Hakuta, 1976). Goldschneider and DeKeyser (2001) use oral-production accuracy as a proxy for acquisition order, relating scores for six grammatical forms across twelve studies to five linguistic and input-related predictors. However, Murakami (2014) questions the small number of morphemes relative to predictors and obtains a similar fit with unrelated predictors. Apparent orders also depend on item selection, the acquisition criterion, and whether individual trajectories or group averages are compared (Pallotti and Peloso, 2008).

3.3 Multiple discovery

Multiple discovery has long been connected to shared intellectual and material conditions. Ogburn and Thomas (1922) emphasize prerequisites and recognition of a worthwhile problem.

Simonton (1978) fits Poisson models to historical counts of multiples, arguing that repeated discovery need not imply inevitability. Brannigan and Wanner (1983) question the equivalence and independence of recorded discoveries and propose a communication model in which announcing a result discourages duplication. Simonton (1986) produces similar distributions of multiplicity under different assumptions about communication and success, and explains how probabilistic discovery can accommodate prerequisites that constrain the order.

A catalog of celebrated doubles omits unsuccessful and unattempted problems. Human discovery races also involve selective effort, as in structural biology (Hill and Stein, 2025b). A fixed machine evaluation can assign every model every problem and retain failures. Independent allocation then predicts overlap ab/M when two systems solve a and b of M problems; shared difficulty can raise that overlap.

Historical accounts also distinguish targets from results, and full from partial contributions (Lynch and Coulter, 2025; Whitty, 2017); peer judgment and reception enter the identification of a discovery (Merton, 1988).

3.4 Shared outcomes, routes, and information

Evolutionary biology separates a shared outcome from a shared route. Shared ancestry, repeated use of related mechanisms, and different mechanisms under similar constraints are distinct (Woods et al., 2006; Stayton, 2015), and replay experiments separate common outcomes from common

sequences (Blount et al., 2008, 2018). The grain of comparison matters (Orgogozo, 2015), as does the difficulty of reconstructing history from endpoints (Losos and Sridhar, 2020). We accordingly compare the ETH papers at the levels of target, theorem, framework, and mathematical engine (Appendix A).

Connected learners also exchange information and methods, which can correlate both what agents know and what they investigate (Zollman, 2007; Weisberg and Muldoon, 2009).

4 Machine multiples: the motivating observation

Different labs have reported machine solutions to the same open problems in nearby periods. The claimed refutations of the Exponential Time Hypothesis (ETH) are the most closely documented pair, and they prompted this paper. They are motivation only: nothing in this section enters the assessment of either form.

Both labs date a discovery milestone to October 16, 2026 (Table 2). The dates are self-reports of different events, a natural-language disproof and a core algorithm, and we did not locate either lab's original artifacts in the public record checked through November 9. The released papers state counting algorithms with the same dependence on the number of variables in their sharp time bounds, and they share a framework for reaching those bounds. They differ in their recursions, parameter dependence, space claims and formalized scope. Appendix A gives the comparison level by level, with three further candidate doubles and a case in which a rumor redirected effort.

Table 2: Reported and recorded dates for the two claimed ETH refutations, all in 2026. Milestones are self-reports with no stated common time zone. Commit times are those recorded in the two repositories and do not establish first public availability (OpenAI, 2026a; Claude and Anthropic, 2026; OpenAI, 2026c; Anthropic, 2026b).

	OpenAI	Anthropic
Reported milestone	Natural-language disproof, October 16; formalization, October 17	Core algorithm, October 16
Date on the paper	October 17, revised October 25; the October 17 text is not public	October 26
Release commit (UTC−7)	October 26, 13:35	October 26, 19:45
Citation of the other	None in the version inspected	Cites OpenAI's preprint

Three features keep the double out of the evidence. It was selected on its outcome: it drew attention because the dates matched. The public releases were not independent events, since Anthropic's release commit follows OpenAI's by about six hours and its paper cites OpenAI's preprint; this dates an opportunity to revise and says nothing about October 16. And a double on the same day asks more than the weak form supplies (Section 2.4): the two lineages must have been at nearby stages, both must have attempted the problem at nearby times, and the problem must have come within reach of both recently, when it might instead have been reachable for some time and only then attempted. Section 7 takes up those alternatives.

5 Evidence from language models

Released OEIS Open evaluations allow a comparison of models attempting the same open problems under a common procedure. Their solved sets are nearly nested across labs and short release ladders. This is strong evidence for nested reach, the consequence of the weak form that a cross-section can show (Proposition 1). Agreement about the order of newly observed solves, which is what the weak form adds, is less well established, and the strong form has little support.

5.1 Shared reach and order on open problems

OEIS Open contains 492 conjectures on 444 integer sequences, with a random subset of 100 forming Lite (Adamczewski, 2026). Our account of their selection and formalization in Tsoukalas et al. (2026) follows Adamczewski. Evaluations use a common networkless agent with Lean and mathematical tools, at $50 per conjecture on the full set and $200 on Lite. We analyzed upstream commit `fd09021e` under a plan committed before per-conjecture verdicts were opened (Appendix B). The plan's order gate required a one-sided 97.5% upper bound on the reversal rate of at most .10 and a lower bound on coverage of at least .20. A comparison meeting both is *supported*; one whose lower reversal bound exceeds .10 is *contradicted*; any other is *inconclusive*.

Nested reach. The plan's primary analysis measures shared reach on the full set: overlap between solved sets beyond what a baseline predicts. Nesting is the particular form of overlap that Proposition 1 requires. Its three models—Opus 4.8, GPT-5.5, and Gemini 3.5 Flash—share the original verifier and evaluation setting. They solve 147, 129, and 109 conjectures, respectively. All 109 Flash solves occur among the Opus solves, as do 126 of GPT-5.5's 129. Across these models and the additional Gemini 3.1 Pro run, 342 conjectures are solved by none, 90 by all four, and 21 by exactly one. Of the six pairs among these four models, two are exactly nested and three have at most three conjectures on the smaller side. The exception is Gemini 3.5 Flash against Gemini 3.1 Pro, the run with a different verifier: each solves conjectures the other misses, 17 and 14.

Table 3 compares overlap with two baselines. All three primary pairs exceed independent allocation at fixed solved counts after Holm adjustment. The second baseline predicts verdicts from statement and description length, proposal year, proposer, selected words, and conjectures per sequence. Adding another model's verdict improves held-out log loss by .47–.55 bits per conjecture across six directions, with all intervals above zero. Folds and resampling are grouped by OEIS sequence.

The overlap thus exceeds what the *prespecified metadata* predict. Those features are not a measurement of mathematical difficulty, which could carry the additional predictive information. Independent allocation is also a weak baseline when problems differ in difficulty. Assigning every model the same list controls target choice within the evaluation; selection of the bank and prior exposure remain possible sources of agreement.

Changing the agent changes one to six of a model's 100 Lite verdicts. Cross-lab differences under a fixed agent range from two to twelve; Opus–GPT-5.5 differs on two to four. Without repetitions, scaffold effects and sampling variation are confounded. Stratifying overlap by how many of the other two full-set models solved each problem gives an odds ratio of 38 for Opus–GPT-5.5 and 3.44 for Flash–Pro, with interval 1.00–10.32 for the latter. Most other intervals are unestimable, and Pro uses a different verifier.

A post hoc audit of accepted proof summaries by the author, without inspecting Lean developments, marked 21 statements as flawed and seven as suspect, all solved by at least three models. Some of the shared reach is therefore shared success on flawed statements. Removing the 21 leaves

Table 3: Overlap on OEIS Open at $50, base agent, 492 conjectures. Δ is in bits per conjecture, for the first model given the second and the reverse.

Pair (solved by each)	Overlap, expected	Lift	p (Holm)	Δ	Δ, reverse
Opus 4.8 – GPT-5.5 (147, 129)	126, 38.5	3.27	$< 10^{-4}$	.544 (.461, .628)	.553 (.470, .637)
Opus 4.8 – Gemini 3.5 Flash (147, 109)	109, 32.6	3.35	$< 10^{-4}$	.475 (.397, .555)	.479 (.397, .566)
GPT-5.5 – Gemini 3.5 Flash (129, 109)	107, 28.6	3.74	$< 10^{-4}$	.527 (.449, .610)	.527 (.449, .611)
Without 21 statements marked flawed (post hoc, $n = 471$):					
Opus 4.8 – GPT-5.5 (126, 108)	105, 28.9	3.63		.496 (.404, .581)	.500 (.408, .586)
Opus 4.8 – Gemini 3.5 Flash (126, 88)	88, 23.5	3.74		.423 (.339, .508)	.422 (.337, .502)
GPT-5.5 – Gemini 3.5 Flash (108, 88)	86, 20.2	4.26		.467 (.380, .550)	.463 (.378, .548)
Two release ladders on Lite ($n = 100$), 97.5% bounds: reversal .005 (.000, .013), coverage .606 (.556, .650); with stages permuted, .500 and .416.					

the overlap and prespecified ordering results unchanged; it does not resolve the distinction between reach and interior acquisition. Solve-cost rank correlations among jointly solved items remain .78–.89.

Order among new solves. The ordering comparison uses two three-release ladders on Lite: Claude Opus 4.8, Fable 5, and Fable 5.1; and GPT-5.5, GPT-5.6 Sol, and GPT-6 Astra. Figure 1 shows the first observed solves. The prespecified analysis assigns each conjecture its first solving rung, censoring those unsolved above the ladder. Solved sets are perfectly nested within each ladder. Across ladders, 2,985 pairs are concordant and 14 reversed: reversal .005, with bounds .000–.013, and coverage .606, with bounds .556–.650. These one-sided 97.5% bounds pass the gate.

The gate could not tell nested reach from interior order. Of the 2,985 concordances, 2,451 involve a conjecture unsolved in one or both ladders. A post hoc comparison restricted to the 53 problems solved in both ladders gives 534 concordant pairs and 14 reversals, a reversal rate of .026, and 504 of those concordances compare the 36 problems both initial releases already solve with the 14 both later solve. The passing result is therefore evidence for nested reach, and it is nearly silent on the order that the weak form adds.

A post hoc decomposition makes the distinction explicit (Table 4). Among the 61 conjectures both initial releases miss, later Anthropic releases solve 14 and later OpenAI releases solve 18, with all 14 shared; independent allocation at those margins predicts overlap of 4.13. Keeping all 61 gives 634 concordances and nine reversals, again including unresolved-at-endpoint problems. Among only the 14 later solved by both, there are 12 concordances and nine reversals: reversal .429 and coverage $21/91 = .231$. Removing only the joint baseline group from the 53 jointly solved problems leaves 17 problems, 30 concordances and 14 reversals.

The shared reach of later releases is clearer than the order of their new solves. Because the restricted cohorts were selected after observing outcomes, including endpoint success, their dependent pair counts diagnose the aggregate agreement rather than supply independent confirmatory evidence: the 14-problem result, with $\gamma = .14$ on 21 resolved pairs, neither provides a well-powered rejection of the weak form nor supports a common fine order. The acquisition interpretation is limited further by single-attempt verdicts on three-rung ladders with a different verifier at the top. With

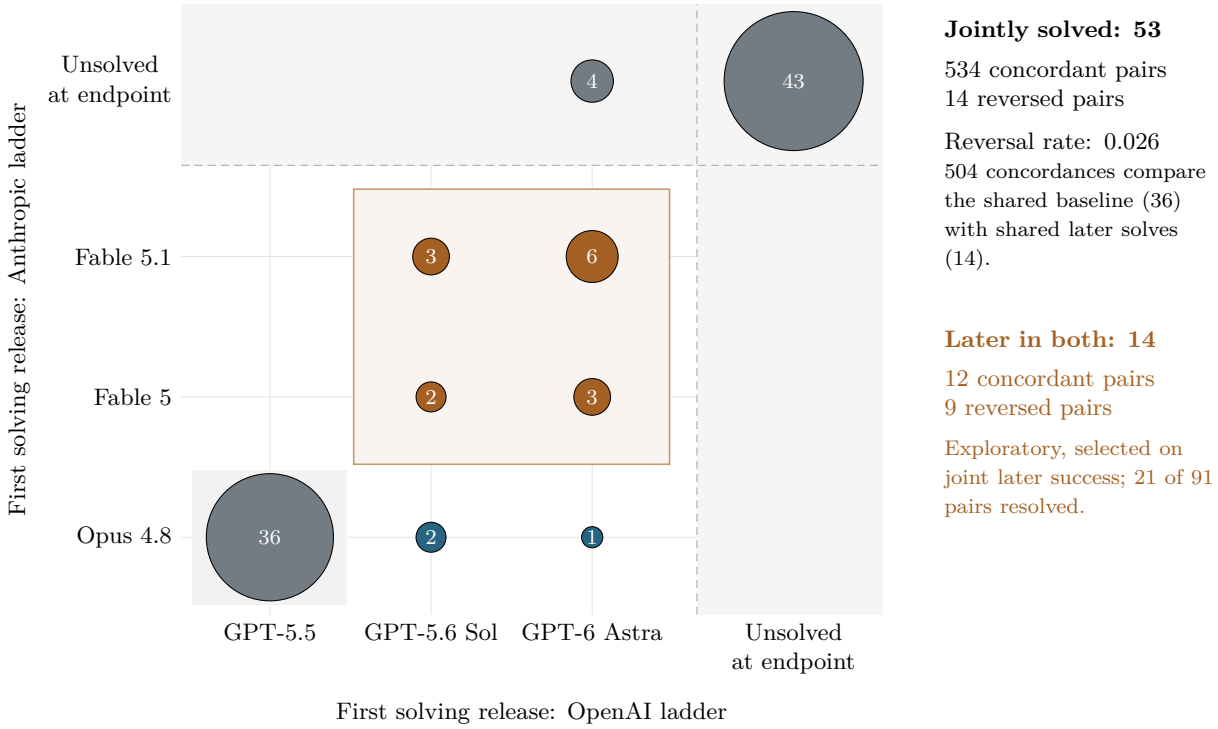


Figure 1: First observed solves on OEIS Open Lite at $200 per conjecture. Bubble areas and labels give conjecture counts; blank cells are zero. Gray bands retain conjectures unsolved at a ladder's endpoint. The jointly solved comparison is post hoc and dominated by baseline-to-later comparisons; Table 4 separates these from order among new solves. Each release–conjecture verdict comes from one attempt.

Table 4: OEIS Lite order decomposed by observation window. Rungs are first observed solves, with one attempt per release–problem cell. The first row reproduces the prespecified all-item count; the cohort decompositions are post hoc. Reversal and coverage are percentages, without new confidence intervals.

Population	n	Concordant	Reversed	Rev.	Cov.
All problems (prespecified test)	100	2,985	14	0.5	60.6
Solved in both ladders	53	534	14	2.6	39.8
Previous row, excluding joint baseline	17	30	14	31.8	32.4
Missed by both initial releases	61	634	9	1.4	35.1
Previous row, later solved by both	14	12	9	42.9	23.1

Coverage uses all unordered pairs within each row's population. Endpoint-selected rows condition on observed success; their pair counts are dependent and do not supply new confirmatory rejection tests. Of the 534 concordances in the jointly solved population, $36 \times 14 = 504$ compare joint baseline solves with joint later solves.

no recorded losses, first and stable crossings coincide in these binary observations, but they do not identify acquisition at probability .5, nor does public release order establish a shared training genealogy, corpus or representation.

The strong form. A post hoc scale check uses the 29 Lite items distinguishing the nine models. On 22, the solvers are exactly the strongest models ranked by total solves; Loevinger's H, which is 1 for perfectly nested patterns, is .87 on the 29 discriminating items (.99 on all 100). On the full set with four models, H is .18 on the 60 discriminating items (.95 on all 492). The populations, model ranges and budgets differ, so these values are not a controlled contrast; retaining the universally solved and universally unsolved items greatly raises both all-item summaries. Of the 22 Guttman errors on the full set, 14 come from the Flash–Pro pair noted above, so the low value mixes scale misfit with a difference of verifier.

The verdict matrices are compatible with a conditional Rasch null that fixes model and item margins. This null assumes independent responses with $p_{mi} = \text{logistic}(\theta_m - b_i)$, a narrower structure than the general increasing function of the strong form. Its rejection could reflect Rasch misfit or dependence without rejecting the strong form; nonrejection does not identify a common scale. In the 10,000 saved Lite null draws, 30 of 36 model-pair co-solve 95% intervals have identical endpoints. The synthetic controls establish neither power at these margins nor equivalence to the null. The result supplies little affirmative evidence for the strong form.

The bank. A Gemini model selected the source conjectures and an agent formalized them. Adamczewski reports that 37% were proposed by Zhi-Wei Sun and 47% have no OEIS links or references. These figures show concentrated authorship and limited linked literature, but they do not directly measure mathematical difficulty, prior exposure, nor research attention. Equal dollar budgets also leave differences in tokens and inference procedures.

5.2 Reachability under changes of method

Studies on known-answer mathematics suggest that post-training changes how readily models reach solutions already available under more extensive sampling. Yue et al. (2025) compare base and reinforcement-trained models at large sample budgets. On AIME24 with 1,024 samples per problem, 63.3% of problems are solved by both, 13.3% by the base alone, and none by the reinforcement-trained model alone. On MATH500 with 128 samples, the corresponding proportions are 92.4%, 3.6%, and

1.0%. Comparisons across six reinforcement-learning algorithms show relatively little variation in their reported sampling-efficiency measure. Short opening cues can also change reliability: prefilling a cue found in training can recover substantial mathematical performance from a base model, and a controlled data intervention changes which cue elicits it (Wang et al., 2026b).

Nesting also appears in published within-lab comparisons on open problems: the novel solutions of GPT-5.4 Pro are included among those of GPT-5.6 Sol Max (Wang et al., 2026a). In the OEIS Open report, 35 of Fable 5's 44 Lite solves were also solved by Opus 4.8 on the full set at a quarter of the budget (Adamczewski, 2026). First Proof systems sharing a base model passed largely the same research-level problems, although those problems had known answers and the comparison had one run per cell (Abouzaid et al., 2026). In comparisons across families, development stage can matter as much as lineage: in our calculations from Liu et al. (2026b), cross-family order correlations for models near in size exceed those of some same-family pairs far apart in size. Similarity peaking near matched performance also appears in Choshen et al. (2022). With corpus and tokenizer fixed, Michaelov et al. (2025) report word-prediction correlations of at least .93 across Pythia, Mamba, and RWKV.

5.3 Variation that a common order must accommodate

OLMo 2's mid-training stage raises GSM8K accuracy from 24.1 to 67.5 while MMLU moves from 59.8 to 63.7 (Team OLMo et al., 2025). It changes the data mixture and includes GSM8K training examples. Increasing code content likewise affects code and natural-language tasks differently (Aryabumi et al., 2024; Liu et al., 2026a). Suppressing syntactic attention structure in BERT changes the appearance of later capabilities (Chen et al., 2023a), while instruction tuning and post-training have different effects across skill factors, mathematical skills, and pretraining durations (Burnell et al., 2023; Mishra et al., 2025; Springer et al., 2025). Even within a fixed recipe, inter-seed agreement in PolyPythias settles near $\kappa = .5$ (van der Wal et al., 2025); word and grammatical development also vary across seeds (Chang et al., 2024; Evanson et al., 2023). Among 100 BERT seeds with similar development-set accuracy, McCoy et al. (2020) find a much larger spread out of distribution. These aggregate and skill-level findings motivate item-level comparisons, with seed variation as a necessary reference for assessing method differences.

The abstract of Lee et al. (2026) reports differential item functioning between Llama and Qwen after conditioning on total score. Smaller models also outperform larger ones on some BIG-Bench tasks (Wei et al., 2022). Such departures count against the strong form; their bearing on the weak form depends on whether they reorder acquisitions. They coexist with broader regularities: Kipnis et al. (2025) report a dominant ability factor, while item-response models predict mathematical benchmark outcomes well in Castleman et al. (2025). Measurement choices, including exact-match scoring and target length, can themselves produce apparently sharp transitions (Schaeffer et al., 2023).

5.4 What might produce the agreement

Existing designs separate several sources of agreement: seed studies hold a recipe fixed; Michaelov et al. (2025) compare architectures under shared corpus and tokenizer; Magnusson et al. (2025) vary data mixtures; and post-training studies compare objectives within a base-model family. Word acquisition tracks input statistics (Chang and Bergen, 2022), while example-learning agreement depends on learner family (Hacohen et al., 2020).

In document search, models agree more with one another than with people at matched accuracy (Borchmann et al., 2026). First Proof's referees also describe better performance on problems

Table 5: Candidate data for further language-model comparisons. Licenses describe the recorded releases and require confirmation before reuse. Only the OEIS analysis was performed here.

Release	License	Possible comparison
LeanOpenProblems-results	none declared	Per-conjecture overlap across labs at equal budget; nesting across generations
DataDecide eval instances	ODC-BY	Is item order more sensitive to corpus than to seed? 25 corpora, fixed recipe; cloze, little math
MathCAMPS responses	MIT	Rank agreement over items across labs against a difficulty null; K–8
OLMo 3 and OLMo 2 variants	Apache-2.0, ODC-BY	With pretraining fixed, does a change of mix or objective reorder items? Needs a small inference run
ElementalTask results	MIT	Rerun Liu et al.'s second hypothesis against a difficulty-only null; four models only
MMLU-Pro eval results	Apache-2.0	Post-training contrast on items; multiple choice
Open LLM Leaderboard details	unstated	Differential item functioning with real power; many derivatives
PolyPythias	Apache-2.0	The seed floor for an item order at 410M; no item outputs
LLM360	Apache-2.0	Order at matched StarCoder exposure

resembling known results (Abouzaid et al., 2026). These observations suggest corpus proximity as one contributor to machine accessibility.

5.5 How far training inputs are shared

The public training-data summaries name Common Crawl for GPT-6 Astra and Claude Opus 5.5 (OpenAI Ireland Ltd, 2026; Anthropic Ireland, Limited, 2026); Gemini 3 Pro's names no particular third-party web corpus (Google, 2026). These disclosures identify sources rather than example-level overlap.

Baack documents common web sources in GPT-3 and released models (Baack, 2024; Baack and Mozilla Insights, 2024); Longpre et al. (2024) examine source-domain overlap. Filtering, partner and user data, and synthetic material introduce further variation. Synthetic data also connects families: OLMo 2's mathematical mixture includes GPT-4o and Qwen2.5 outputs (Team OLMo et al., 2025).

5.6 Tests the available data permit

Table 5 lists data for further comparisons. OLMo variants could test whether changing the objective changes acquisition order beyond seed variation; DataDecide could compare corpus and seed effects. The prospective chain test in Section 8.1 follows acquisition on a fixed bank with repeated attempts.

6 A verifiable precedent: chess

Chess permits repeated evaluation across many training checkpoints at modest cost. We measure shared accessibility and acquisition order on first-move decisions from historical self-play runs. Their ancestry and data independence are unestablished. Self-play runs generate their own training data, so chess cannot test the clause on shared information: agreement here speaks to problem structure and learner family at least as much as to a common corpus. What these comparisons provide is a provisional precedent for the weak form, a test of whether an acquisition order transfers at all, at a resolution the language-model data lack.

6.1 Existing evidence

On Lichess, Zahavy et al. (2023) report about 84.94% solved by one AlphaZero seed and 93.10% by the union of three, read from their figure. Shared failures exceed the expectation under homogeneous, independent failures, as a shared difficulty scale would predict. Starting self-play from puzzle positions also changes performance substantially. Other endpoint comparisons examine learned piece values (Tomašev et al., 2020; Pálsson and Björnsson, 2023), agreement with human moves (Silver et al., 2017; McIlroy-Young et al., 2020), shared blind spots (Wang et al., 2023), or the prediction of human difficulty (Björkqvist, 2025).

Published developmental studies examine the acquisition of a small set of chess concepts (McGrath et al., 2022), while work on Hex and Go shows that learned skills need not appear in a simple monotone sequence (Lovering et al., 2022; Tian et al., 2019). Related measurements of depth, player strength, human error, and item difficulty provide methods for comparison (Biswas and Regan, 2015; Regan and Haworth, 2011; Anderson et al., 2016; Martínez-Plumed and Hernández-Orallo, 2018).

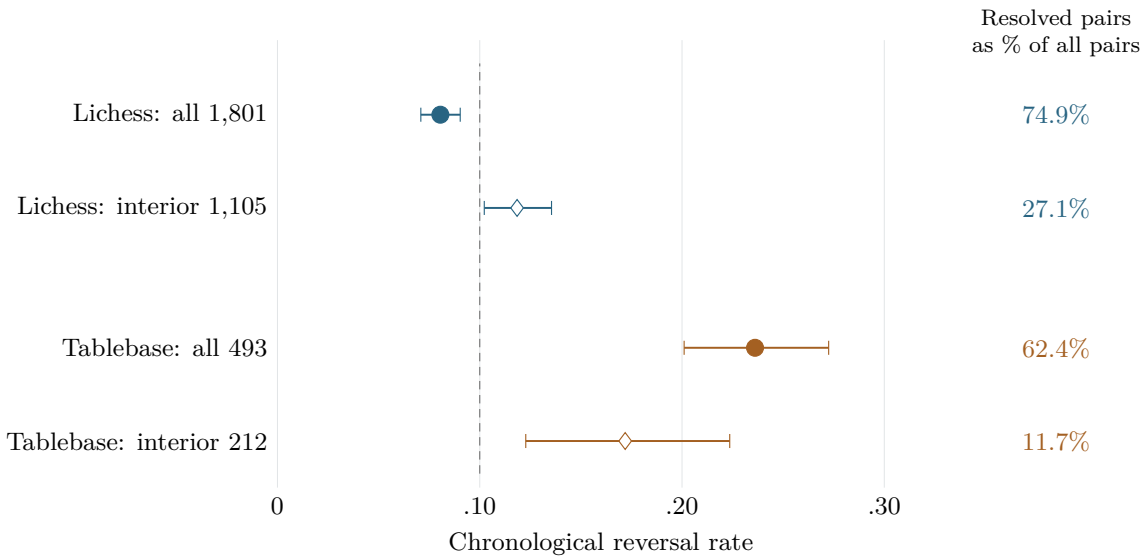
6.2 Design in brief

We evaluated 24 checkpoints from each of eleven Leela runs, naming T76 and T77 as the primary pair before opening the outcomes and using the other nine as comparisons against each reference. The resulting 18 comparisons share items and reference runs; they are not independent replications. The banks contain Lichess puzzles presented with game history and tablebase endgames presented as bare positions. The primary budget is one visit, which evaluates the policy's preferred first move; 16 visits provide a search-budget check for the primary pair. Success means an accepted first move. The endpoint does not require completing the solution line, and search may still have influenced the policy's training targets. A preceding Stockfish pilot tested the measurement procedure. Appendix C gives the implementation, validation, and deviations.

Accessibility integrates each item's solved indicator over a common strength range defined by separate anchors, while the chronological endpoint compares stable acquisition through the sampled window using censored intervals. The latter identifies the final uninterrupted sequence of solves through the observation endpoint and is the closest measurement here to the weak form; under the deterministic evaluation procedure, solve probabilities are zero or one. Accessibility orders the same solves on a strength axis in place of the training clock; under the item side of the strong form it too should agree across runs. The plan required support on both banks for a broad claim. A strength-matched comparison is *descriptive only* when too few checkpoints span the common strength range; chronological comparisons do not use that range. Bounds resample items with their complete recorded trajectories to retain dependence among derived pairs, conditioning on checkpoint histories and, for predictive intervals, fitted predictors. Anchor composition is examined separately; these calculations do not estimate variability over newly trained lineages.

6.3 Agreement and failures

On Lichess, T76 and T77 meet the accessibility gate at both budgets and the chronological gate. At one visit, accessibility reversal is .034, with bounds .029–.040 and coverage .730. Chronological reversal is .081, with bounds .071–.090 and coverage .749. Another run's accessibility improves prediction beyond item features and Stockfish difficulty by about .3 bits in each direction. At matched strength, overlap in items newly solved on the five-checkpoint grid exceeds the specified difficulty-model expectation in all four windows. This first-solve endpoint differs from stable acquisition. Among the 18 replication comparisons on Lichess, 15 accessibility comparisons are supported and three inconclusive; none is contradicted.



Filled circles: prespecified all-item populations. Open diamonds: exploratory uncensored-in-both subgroups.
Bounds condition on recorded histories; subgroup bounds also condition on outcome-based selection.

Figure 2: Chronological reversals for the complete primary populations and the post hoc interior subsets. Complete-population bounds are the prespecified item-bootstrap bounds; interior bounds come from new paired item resampling conditional on fixed, outcome-selected subgroups. Both use 2.5th and 97.5th percentile endpoints. Coverage uses all original item pairs in each bank. The subgroup bounds do not adjust for selection or quantify uncertainty over new training lineages; no new confirmatory status is assigned.

A post hoc decomposition separates acquisition observed inside both windows from order constrained partly by censoring (Figure 2; Table 14). Among the 1,105 Lichess items with uncensored acquisition in both runs, 52,002 of 438,976 resolved pairs reverse: .118, compared with .081 in the complete analysis. The interior rate corresponds to $\gamma = .76$. It lies above the .10 threshold set for the whole-bank test and far below the .5 of unrelated orders. These pairs cover .271 of all original item pairs and constitute .362 of the complete analysis's resolved pairs. Thus, the full result combines agreement within the observed acquisition window with stronger agreement on pairs involving already-acquired or still-unacquired items. The subset is selected from these trajectories and receives no new confirmatory status; censored comparisons remain valid order evidence in the original analysis.

Learning is also nonmonotone at the sampled checkpoints: 815 of 1,801 Lichess items in T76 and 785 in T77 have at least one observed solve-to-fail transition. Stable acquisition here means retention through a fixed observed endpoint, not permanent acquisition.

The tablebase bank gives a different result. The primary strength-matched comparison lacks that range and is descriptive only at one visit. Its chronological comparison is contradicted, with reversal .236 and bounds .201–.272. The prespecified 16-visit strength comparison is likewise descriptive only; a post hoc comparison of the two 16-visit curves on their own common range is contradicted. Thus, neither budget supports the broad two-bank claim. By Proposition 2, a contradiction of the weak form is also one of the strong.

The accessibility comparisons (Figure 3, Appendix C) show a format split: tablebase reversals

range from .054–.136 within formats and .226–.271 across them. Lichess shows no comparable separation. Within the classical format, T40–T76 and T50–T76 have contradicted accessibility comparisons at .136 and .132; their chronological comparisons are inconclusive at .106 and .101. T59–T76 and T70–T76 have inconclusive accessibility comparisons near .11. The within-format results therefore challenge a common accessibility ranking without establishing a chronological reversal rate above .10. Table 13 in Appendix C gives the remaining endpoints and sensitivities.

The primary tablebase contradiction compares runs with different input formats. The clause on representation in the hypothesis predates this result, and it places such a comparison at the edge of the hypothesis's scope. The clause does not dispose of the failure. Within a format, tablebase chronology is inconclusive and two accessibility comparisons are contradicted, and the same change of format moves nothing on Lichess.

The largest tablebase gap occurs on pawnless positions, where canonical formats apply larger symmetry transformations. This post hoc association is confounded with training era and recipe.

A follow-up with readings fixed before its outcomes removed game history from Lichess and disabled history-plane filling for tablebase positions. On 697 Lichess items matching the tablebase presentation convention, the format gaps were .005 on accessibility and .006 chronologically, against tablebase baselines of .170 and .153. All seven bare-Lichess pairs met the accessibility gate; four chronological comparisons remained inconclusive.

Blank history planes impaired classical networks enough to remove their strength axes: T75's anchor success fell from .32 to .25 across training, against a final .77 in the original condition. The tablebase arm of the follow-up was therefore partial. Only two Lichess items are pawnless, leaving presentation on that population unresolved. A symmetric-copies check remains deferred.

7 Rival explanations

A double can arise when two labs attempt a problem their models could already solve. The timing then reflects opportunity rather than a recent acquisition.

We distinguish three questions: does an order transfer, why would it transfer, and why did a particular double happen? Capability growth on a common scale belongs within the strong form; target selection may explain a double's timing alongside it. Table 6 sets out the comparisons.

7.1 Target selection and opportunities to attempt

Attention can determine when a capable system first attempts a problem. OpenAI describes a sweep of approximately 4,000 research problems, but the record we examined identifies neither ETH's inclusion nor its first attempt (OpenAI, 2026c). Anthropic's provenance likewise leaves the reason for its start date unstated (Claude and Anthropic, 2026).

The October 5 release of the 3SUM/APSP result and OpenAI's October 6 mathematical collection preceded the reported ETH milestones by eleven and ten days. Progress in adjacent complexity problems could have encouraged an ETH attempt, although neither paper attributes its choice to these events and Anthropic already knew its 3SUM result (Alman and Vassilevska Williams, 2026; OpenAI, 2026e; Claude and Anthropic, 2026).

Two Anthropic results emerged during work on other tasks (Alpöge and Furman, 2026; Alman and Vassilevska Williams, 2026). Selection can nevertheless operate at the level of a research area, as effort also concentrates in human discovery races (Hill and Stein, 2025a).

FrontierMath Erdős illustrates heterogeneous reach within one model: #74 was solved in six of six attempts and #126 in four of four, while 56 conjectures were unsolved in 172 attempts

(Adamczewski and Bloom, 2026). Dated prompts, earlier failures, and fixed evaluation schedules could distinguish the timing of attempts from acquisition.

7.2 Information transfer

Information can pass through training data, runtime tools, target selection, and revision of a paper. The ETH accounts leave runtime information access unspecified. Another Anthropic project describes a network-enabled literature agent (Alpöge and Furman, 2026); its use for ETH is unknown. The Navier–Stokes episode supplies a reported example of rumors redirecting effort (Section A.5).

Networkless evaluation prevents runtime retrieval from the external network but leaves training exposure to published solutions possible. Access to local tools, libraries and supplied context also belongs in the information-access specification. The OEIS repository contains accepted proofs, motivating the held-out-bank and information-access requirements in Section 8.

7.3 Capability and the timing of discoveries

Capability growth could align discovery dates if lineages reach nearby acquisition stages at nearby times and attempt the same problems (Section 9.2). The suggestion that both ETH systems crossed the same threshold has appeared informally (Brooks, 2026), but their parity remains unmeasured. In FrontierMath Erdős, one model solved two problems while four others solved none (Adamczewski and Bloom, 2026); predictable progress and a common path are also distinct in other domains (Kryazhimskiy et al., 2014). Comparable earlier failures followed by reliable success after a model change would support a capability transition, but the ETH records inspected here lack that controlled before-and-after comparison.

7.4 Human ripeness, chance, and exposition

Common mathematical ingredients can make a problem ripe for several discoverers; for ETH, the question is how naturally the shared framework follows from those ingredients. Carenini connects the random k-SAT work to a common preceding report and earlier threshold results, including Abbe and Montanari (2014), known here through that account (Carenini, 2026). An unlocking idea can synchronize work (Whitty, 2017). Assessing chance also requires a defined population and time window: independent uniform dates within W days would match with probability $1/W$, but neither assumption is justified for ETH, which attracted attention because its reported dates matched. Anthropic's citation of OpenAI's preprint creates an opportunity for its later presentation to respond to that work.

7.5 Why an order would transfer

A corpus account predicts transferable accessibility from similar relevant information. Changing problem-specific training information while holding the rest fixed would test this explanation through changes in acquisition order. Hint length offers another measure if missing knowledge can be separated from elicitation (Section 2.5). Contamination predicts agreement from exposure to the evaluated solutions; comparing public and unpublished banks requires establishing exposure and controlling composition and difficulty.

A representational-convergence account (Huh et al., 2024) requires similarity in geometry to extend to acquisition sequence. A problem-structure account instead locates the common constraints in the tasks. Neither entails invariance to representation changes or agreement with human difficulty without further assumptions. Shared practices provide another possibility: studies report similarity

Table 6: Three explanatory questions and informative comparisons.

Question	Comparison	Alternatives or conditions
Does an order transfer?	Repeated acquisition measurements, with censored and within-window contributions separated	Lineage-specific orders; boundary separation; metadata and unmeasured difficulty
Why would it transfer?	Controlled changes to corpus, representation, and training; public versus unpublished banks	Corpus proximity, contamination, convergent representations, and problem structure
Why did a double happen?	Dated attempts, model snapshots, access records, and proof histories	Aligned stages and opportunities; target choice, leakage, ripeness, chance, and write-up

across providers in scientific methods and research outputs (Bisht et al., 2026; Carlon et al., 2026), and increased error agreement after instruction tuning (Goel et al., 2025). The chess agreement in Section 6 arises among self-play runs with no common corpus, which for that domain favors problem structure or learner family over corpus proximity.

8 Prospective tests and forecasts

The evidence motivates three prospective questions: whether acquisition order transfers across documented training trajectories, whether an existing accessibility ranking predicts later performance, and how often another lab's model replicates a new solution. The first tests the weak form directly. The second concerns nested reach, the consequence both forms share. The third concerns the strong form. These are proposed studies, whose populations, sample sizes and numerical decision rules remain to be fixed before target outcomes are examined. Section 8.3 adds the one numerical forecast we commit to now.

8.1 A direct test of acquisition order

Follow several documented model lineages through training on a common held-out problem bank. Repeated attempts at each checkpoint estimate solve probabilities under a fixed prompt, tool allowance, budget and acceptance rule. Fix the observation endpoint in advance and measure stable acquisition as defined in Section 2.1. Include multiple seeds to distinguish ordinary run-to-run variation from differences associated with training method.

The central comparison is the sequence of acquisition across learners receiving comparable information. Different architectures, sizes and optimization procedures are candidate sources of variation; the eligible differences and the meaning of shared inputs must be specified before evaluation. A same-architecture recipe comparison could first establish the measurement's resolution. Opaque commercial release ladders provide observational comparisons, with less control over the information available to each learner.

The study should report both absolute order agreement and how much between-method disagreement exceeds within-method seed variation. It must also separate order involving window boundaries from order among acquisitions observed inside both windows. A persistent division into already-solved and never-solved problems supplies little evidence about the latter, even when its aggregate ordering score is excellent.

Uncertainty around the acquisition threshold affects both pair order and membership in the

interior population. Retain that uncertainty rather than dropping ambiguous problems, and account for the dependence of pairs sharing a problem. Fix minimum temporal coverage and the treatment of missing observations and censoring in advance. First crossing and thresholds .25 and .75 provide sensitivities to the primary .5 acquisition definition. They also separate the forms. By Proposition 2 the strong form requires one order at every threshold, so an order that changes with the threshold is compatible with the weak form and counts against the strong. Our present data cannot make this comparison, since both sets record binary outcomes.

Before choosing sample sizes and checkpoint spacing, assess the procedure on simulated shared orders, different learning speeds, temporary losses and meaningful reversals. Repeated-attempt measurements on development tasks can establish feasible budgets and attainable resolution. A candidate design offers one conservative uncertainty construction, but its population and numerical choices have not been adopted (Appendix B). Reproducible reversals in an adequately observed, prespecified population would count against the weak form and, with it, the strong; agreement confined to the window boundaries would leave the weak form's temporal component unresolved.

8.2 A forecast of accessibility transfer

The four full-set models evaluated with the base agent at $50 rank the 492 OEIS Open conjectures by the number solving each one (Adamczewski, 2026). The proposed forecast is that this ranking transfers to later models: among pairs with unequal reference counts, solving only the higher-ranked item should be more common than solving only the lower-ranked one. Report the reversal rate and the fraction of eligible pairs resolved; solving both or neither leaves a pair unresolved.

This is a forecast of existing accessibility, and so of nested reach. The zero-to-four reference solve-count strata are $[342, 21, 7, 32, 90]$. Of 57,743 eligible pairs, 51,300 (88.84%) cross the zero/nonzero boundary. A future model solving exactly the existing 150-item union would obtain zero reversals and 88.84% coverage without a new solution. The 342 currently unsolved conjectures remain tied. Comparisons across the zero/nonzero boundary and within the positive-score strata should therefore be shown separately.

Before evaluating the forecast, fix model eligibility, exposure requirements, the harness and budget, uncertainty accounting and decision thresholds. Repeated reversals on a well-resolved bank would challenge transfer of this accessibility ranking, although failure of the proxy would not alone reject the weak form. A model solving everything or nothing supplies no ordering information.

Accepted proofs were public by September 16, 2026, and some may have appeared earlier. A clean item-level evaluation requires documented training freezes predating relevant proof exposure. A fresh unpublished bank offers another route: evaluate reference models, commit their ranking, then observe later models on that same bank. It would be a separate prospective study. Predicting the sequence of discoveries among the current 342 unsolved conjectures would require a new ranking rule.

8.3 A dated forecast of nested reach

The ranking forecast above can be met by existing reach alone. We therefore commit to one prediction restricted to new solves. Take the 43 Lite conjectures unsolved at the endpoint of both ladders in Figure 1. For the first later release from each provider evaluated under the same $200 procedure, let a and b be the numbers of these conjectures that each solves. We forecast that at least three quarters of the smaller set also lies in the larger. Independent allocation predicts a share of $\max(a, b)/43$. The forecast is scored on the first such pair of evaluations reported before 31 December 2027 with $\min(a, b) \geq 4$ and $\max(a, b) \leq 21$, so that the independent share is at most one

half. A conjecture is excluded if an accepted proof of it was public before the stated training cutoff of the release evaluated later. The corresponding figure in the present data is 14 of 14, among the 61 conjectures both initial releases missed. This forecast is dated 13 November 2026.

Solved sets that cross would count against nested reach and so, by Propositions 1 and 2, against both forms. Success would support the consequence the forms share and leave interior order where it stands.

8.4 A forecast of conditional replication

When one lab's model solves a recognized open problem, how often can another lab's comparably capable model solve it under a fixed procedure? The proposed forecast is a substantial conditional replication rate. Its numerical target remains open alongside the problem population and capability criterion. The model side of the strong form motivates the comparison but supplies no universal rate (Section 2.4).

The motivating OEIS observation concerns the 61 Lite conjectures both initial releases missed. Later OpenAI releases solved all 14 later Anthropic solves, against four of the other 47; Anthropic solved 14 of OpenAI's 18, against none of the other 43. FrontierMath Erdős supplies a contrasting observation: one model solved two of 68 problems and four others solved none (Adamczewski and Bloom, 2026). A prospective estimate needs a population and procedure chosen independently of such outcomes.

Freeze a recognized problem list, enrollment window and model-selection rule. Use immutable replicators predating each discovery, with independent capability anchors and no access to the discovered proof. Fix attempts, budgets and proof-acceptance rules; assess mathematical statement fidelity as well as proof validity. Eligibility cannot be revised after a failed replication. Discoveries with no eligible counterpart must be distinguished from attempted replications that fail.

Specify the weighting of problems and models, minimum sample and missing-data treatment before results are opened. Report per-problem outcomes alongside the aggregate conditional rate, retaining singletons and unsuccessful attempts. Frequent failures among eligible counterparts would count against the replication forecast. Explaining synchronized discovery dates additionally requires records of earlier attempts and capability transitions.

8.5 What would change the interpretation

Agreement accounted for by independently measured mathematical difficulty would narrow the need for a specifically machine-based explanation. Controlled changes to corpus, representation and training procedure could distinguish causes of transfer. Granularity and scoring must be fixed, since they can change apparent order (Chen et al., 2023b); additional pathway-predictability measures remain supplementary (Hosseini et al., 2019).

A quantitative test of the strong form would require an independently calibrated capability coordinate, a metric and tolerance for approximation, and held-out probability comparisons against richer alternatives. For ETH, comparable earlier failures followed by reliable success after a model change would support an acquisition-based explanation of timing. High success rates in older models would instead favor attempt timing. Original artifacts and access records are needed to distinguish those accounts.

Table 7: The evidence by claim. Each claim is implied by the one below it. Reversal rates are among resolved pairs; $\gamma = 1 - 2r$.

Claim	Evidence	Standing
Nested reach	OEIS full set: of six model pairs, two are nested, three have at most three conjectures on the smaller side, and one crosses; another model's verdict adds .47–.55 bits beyond metadata. Lite: all 14 later Anthropic solves are among OpenAI's 18	Holds. Unmeasured difficulty, solution exposure and 21 flawed statements qualify what is shared
Weak form: interior order	Lichess: interior reversal .118 ($\gamma = .76$), whole bank .081. Tablebase: .236 across input formats, contradicted; same-format chronology inconclusive. OEIS: 12 concordant and nine reversed among 14 new solves	Partial on chess. Uninformative on open problems. No repeated-attempt trajectories for language models
Strong form	Loevinger's H of .87 on Lite and .18 on the full set; a Rasch null not rejected, with little power; differential item functioning reported elsewhere	Little support. A powered test at the level of probabilities is open

9 Discussion

9.1 What the evidence bears on

We read the evidence against the three claims of Section 2: nested reach, the weak form, and the strong form (Table 7). Nested reach holds on OEIS Open, across labs and within release ladders, beyond what the prespecified metadata predict. The weak form's own content is order among problems acquired inside both observation windows. That order is supported on one chess bank at a rank correlation near .76, contradicted on the other across input formats, and unobserved for language models: the one interior comparison available on open problems has 21 resolved pairs and no repeated attempts. The strong form has little support, and differential item functioning has been reported elsewhere.

The whole-population order tests that passed their gates, on OEIS and on Lichess, drew most of their resolved pairs from comparisons involving a window boundary. The exploratory decompositions measure how much, without replacing the original tests or supplying new confirmatory rejections. Because the strong form implies the weak, the tablebase contradiction counts against both, to the extent that a change of input format lies within the hypothesis's scope. A prospective study must specify temporal evidence and uncertainty treatment before seeing its outcomes.

The weak form makes the narrower commitment to sequence while allowing probability trajectories to differ. The strong form would offer a more economical account if a common scale predicted item probabilities across lineages as well as separate lineage models. The existing Rasch diagnostic and scale summaries leave that comparison unresolved: the former tests a narrower model with limited power, and the full and Lite summaries cover different model sets and budgets.

9.2 From an order to doubles

The weak form contributes to doubles by fixing which problems lie at a lineage's frontier. It leaves open whether two lineages are at the same frontier, and how often a second attempt succeeds (Section 2.4). Equal pace alone is insufficient: lineages can advance equally fast while remaining far

apart. The rate of replication belongs to the strong form, through the spread of solve probabilities across problems, and a numerical replication forecast needs its own empirical justification and a specified population. The ETH double that motivated this paper therefore leans on the form with the least support.

Moving from a fixed conjecture bank to major discoveries requires records of attempts and a consistent account of what constitutes the same result. The ETH comparison illustrates a method for comparing results; a prospective program would also need attempt records. The proposed replication study measures shared reach. Explaining discovery dates additionally requires evidence of transitions near those dates, while replication by an older model could favor an explanation based on when the problem was attempted.

9.3 If an order exists

A transferable order could improve forecasts and guide evaluation. Banks concentrated at one acquisition stage may conceal model differences or cease to discriminate once that stage has passed. The present OEIS ranking leaves the currently unsolved conjectures tied.

Correlated failures can limit gains from combining systems, while departures can supply useful diversity, as the chess unions and OEIS singletons illustrate. Ensemble gains depend on item-level error dependence and cost, which a common acquisition order does not determine. Algorithmic monoculture can reduce aggregate quality (Kleinberg and Raghavan, 2021) and impose epistemic costs on human communities (Fazelpour et al., 2026).

Machine doubles also complicate priority. A result, framework, construction, formal proof, and explanation can be distinct contributions; the ladder of sameness in Appendix A specifies which is compared. Mathematical goals extend to understanding and the choice of worthwhile questions (Tao, 2026).

10 Limitations and disclosure

10.1 Limits of the evidence

Our OEIS release-ladder comparison uses single attempts at three releases per provider, with incomplete training genealogy and different verifiers at the top rungs. Our chess trajectories concern deterministic first-move decisions. Both leave a gap to repeated-attempt acquisition in documented language-model lineages. The tablebase split remains unexplained, and its concentration on pawnless positions awaits a symmetry check. Propositions 1 and 2 relate the exact forms; the approximate forms are connected only through the margin and the two references of Section 2.1.

The .10 reversal and .20 coverage thresholds are practical commitments, not inferred constants. The proposed studies require populations and decision rules supported by feasibility and power assessments. The ETH comparison lacks original discovery artifacts. Reading depth for each source is recorded in the companion repository; the principal secondary and abstract-only uses are identified in the text.

10.2 Disclosure and attribution

Machine assistance on implementation, engine runs, and analysis code was provided by Anthropic's Claude Opus 5.5 and OpenAI's GPT-6 Astra Pro. On revision, a suggestion for the values of the dated forecast in Section 8.3 was accepted from Anthropic's Claude Fable 5.2.

Data and code availability

The companion research materials contain the manuscript sources, analysis code, prespecified protocols, derived results and raw chess records. They support aggregate OEIS checks and raw-record chess reanalysis. Network weights are identified by hashes. The pinned per-conjecture OEIS source is excluded pending clarification of redistribution terms, so the package does not independently verify every upstream verdict. A repository containing these materials will be made public at publication time.

Acknowledgments

We thank DTU HPC for computational resources. Machine assistance is described in Section 10.2.

References

Emmanuel Abbe and Andrea Montanari. On the concentration of the number of solutions of random satisfiability formulas. *Random Structures & Algorithms*, 45(3):362–382, 2014.

Mohammed Abouzaid, Nikhil Srivastava, Rachel Ward, and Lauren Williams. First Proof second batch. arXiv:2606.18119, 2026.

Tom Adamczewski. OEIS Open: How many conjectures can language models turn into theorems? arXiv:2608.11941, 2026.

Tom Adamczewski and Thomas F. Bloom. FrontierMath Erdős. arXiv:2609.25050, 2026.

Josh Alman and Virginia Vassilevska Williams. Truly subquadratic 3SUM and truly subcubic APSP via triangles in sparse lopsided graphs. arXiv:2610.06783, 2026.

Levent Alpöge and Ralph Furman. More than two thirds of the zeros of the Riemann zeta function are simple and on the critical line. arXiv:2608.13637, 2026.

Ashton Anderson, Jon Kleinberg, and Sendhil Mullainathan. Assessing human error against a benchmark of perfection. In *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2016. doi: 10.1145/2939672.2939803.

Anthropic. ETH main results: Formal statements and input encoding. anthropics/formal-math, eth/MainResults.lean, main branch inspected 9 November 2026, 2026a. URL https://github.com/anthropics/formal-math/blob/main/eth/MainResults.lean.

Anthropic. anthropics/formal-math: READMEs for eth, 3sum-apsp and zeta23 at revision 16a5ce4, 2026b. URL https://github.com/anthropics/formal-math.

Anthropic Ireland, Limited. Public summary of training content: Claude Opus 5.5, 2026. URL https://aial.ie/research/gpai-training-transparency/archive/Claude_Opus_5_5_2026_09_28.pdf.

Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code, or not to code? Exploring impact of code in pre-training. arXiv:2408.10914, 2024.

Stefan Baack. A critical analysis of the largest source for generative AI training data: Common Crawl. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 2024. doi: 10.1145/3630106.3659033. URL https://facctconference.org/static/papers24/facct24-148.pdf.

Stefan Baack and Mozilla Insights. Training data for the price of a sandwich: Common Crawl's impact on generative AI. Technical report, Mozilla Foundation, 2024. URL https://foundation.mozilla.org/documents/350/Common_Crawl_Mozilla_Foundation_2024.pdf.

Camilla Bettoni and Bruno Di Biase, editors. *Grammatical Development in Second Languages: Exploring the Boundaries of Processability Theory*, volume 3 of *Eurosla Monographs Series*. European Second Language Association, 2015. URL http://eurosla.org/monographs/EM03/EM03tot.pdf.

Harshit Bisht, Vinay Kumar, Kevin Maik Jablonka, Mausam, and N. M. Anoop Krishnan. Agentic AI scientists are not built for autonomous scientific discovery. In *Advances in Neural Information Processing Systems, Position Paper Track*, 2026.

Tamal T. Biswas and Kenneth W. Regan. Quantifying depth and complexity of thinking and knowledge. In *Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART 2015)*, pages 602–607, 2015. doi: 10.5220/0005288306020607.

Sebastian Björkqvist. Estimating the difficulty of chess puzzles by combining fine-tuned Maia-2 with hand-crafted and engine features. In *Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS)*, volume 43 of *ACSIS*, pages 801–806, 2025. doi: 10.15439/2025F6497.

Zachary D. Blount, Christina Z. Borland, and Richard E. Lenski. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 105(23):7899–7906, 2008. doi: 10.1073/pnas.0803151105.

Zachary D. Blount, Richard E. Lenski, and Jonathan B. Losos. Contingency and determinism in evolution: Replaying life's tape. *Science*, 362(6415):eaam5979, 2018. doi: 10.1126/science.aam5979.

Łukasz Borchmann et al. Strategic navigation or stochastic search? How agents and humans reason over document collections. arXiv:2603.12180, 2026.

Augustine Brannigan and Richard A. Wanner. Historical distributions of multiple discoveries and theories of scientific change. *Social Studies of Science*, 13(3):417–435, 1983. doi: 10.1177/030631283013003004.

Richard Brooks. Two machines broke the same rule, 2026. URL https://richard-brooks.com/two-machines-broke-the-same-rule/.

Tristan Buckmaster. Statement on finite-time blowup results with smooth forcing and dealings with OpenAI. PDF on personal web page, created September 8, 2026, 2026. URL https://cims.nyu.edu/~tristanb/statement.pdf.

Ryan Burnell, Han Hao, Andrew R. A. Conway, and Jose Hernandez Orallo. Revealing the structure of language model capabilities. arXiv:2306.10062, 2023.

Gaia Carenini. A polynomial scaling window for random k-SAT and a proof of the satisfiability conjecture. ECCC TR26-229, 2026. URL https://eccc.weizmann.ac.il/report/2026/229/.

Francesca Carlon, Brecht Verbeken, Vincent Ginis, and Andres Algaba. Thinking like a scientist? A structural study of LLM-generated research methods. arXiv:2606.26130, 2026.

Jane Castleman, Nimra Nadeem, Tanvi Namjoshi, and Lydia T. Liu. Rethinking math benchmarks: Implications for AI in education. In *iRAISE Workshop at AAAI 2025*, volume 273 of *PMLR*, pages 66–82, 2025. URL https://proceedings.mlr.press/v273/castleman25a.html.

Timothy M. Chan and R. Ryan Williams. Deterministic APSP, orthogonal vectors, and more: Quickly derandomizing Razborov–Smolensky. *ACM Transactions on Algorithms*, 17(1):2:1–2:14, 2021. doi: 10.1145/3402926. URL https://people.csail.mit.edu/rrw/derand-apsp-ov-cr-talg.pdf.

Tyler A. Chang and Benjamin K. Bergen. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16, 2022.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. *Transactions of the Association for Computational Linguistics*, 12:1346–1362, 2024.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. arXiv:2309.07311, 2023a.

Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! A data-driven skills framework for understanding and training language models. arXiv:2307.14430, 2023b.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8281–8297, 2022. doi: 10.18653/v1/2022.acl-long.568.

Claude and Anthropic. A $2^{o(n)}$ time algorithm for 3-SAT. Paper in anthropics/formal-math, dated October 26, 2026, 2026. URL https://github.com/anthropics/formal-math/tree/main/eth.

Linnea Evanson, Yair Lakretz, and Jean-Rémi King. Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Sina Fazelpour, Joseph O'Brien, and Hannah Rubin. Navigating epistemic monocultures in AI-driven science: A simulation study. arXiv:2608.19390, 2026.

Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, 2025.

Jennifer M. Goldschneider and Robert M. DeKeyser. Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning*, 51(1): 1–50, 2001. doi: 10.1111/1467-9922.00147. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9922.00147.

Google. Public summary of training content: Gemini 3 Pro family, 2026. URL https://aial.ie/research/gpai-training-transparency/archive/Gemini_2026_07_21.pdf.

Guy Hacohen, Leshem Choshen, and Daphna Weinshall. Let's agree to agree: Neural networks share classification order on real datasets. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, pages 3950–3960, 2020.

Kenji Hakuta. A case study of a Japanese child learning English as a second language. *Language Learning*, 26(2):321–351, 1976. doi: 10.1111/j.1467-1770.1976.tb00280.x.

Ryan Hill and Carolyn Stein. Race to the bottom: Competition and quality in science. *The Quarterly Journal of Economics*, 140(2):1111–1185, 2025a. doi: 10.1093/qje/qjaf010.

Ryan Hill and Carolyn Stein. Scooped! Estimating rewards for priority in science. *Journal of Political Economy*, 133(3):793–845, 2025b. doi: 10.1086/733398.

Sayed-Rzgar Hosseini, Ramon Diaz-Uriarte, Florian Markowetz, and Niko Beerenwinkel. Estimating the predictability of cancer evolution. *Bioinformatics*, 35(14):i389–i397, 2019. doi: 10.1093/bioinformatics/btz332.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, 2024.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, 2025.

Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench – a sparse benchmark of reasoning and knowledge in large language models. In *International Conference on Learning Representations*, 2025.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021. doi: 10.1073/pnas.2018340118.

Stephen D. Krashen. *Principles and Practice in Second Language Acquisition*. Pergamon, Oxford, 1982. URL http://www.sdkrashen.com/content/books/principles_and_practice.pdf.

Sergey Kryazhimskiy, Daniel P. Rice, Elizabeth R. Jerison, and Michael M. Desai. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*, 344(6191):1519–1522, 2014. doi: 10.1126/science.1250939.

Hansol Lee, Jason Cho, Nick Haber, Benjamin Domingue, and Sanmi Koyejo. Differential item functioning as an item-level diagnostic for LLM benchmarks. In *Advances in Neural Information Processing Systems*, 2026. URL https://neurips.cc/virtual/2026/poster/139203.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *International Conference on Learning Representations*, 2016.

Emmy Liu, Amanda Bertsch, Lintang Sutawika, Lindia Tjuatja, Patrick Fernandes, Lara Marinov, Michael Chen, Shreya Singhal, Carolin Lawrence, Aditi Raghunathan, Kiril Gashteovski, and Graham Neubig. Not-just-scaling laws: Towards a better understanding of the downstream impact of language model design decisions. arXiv:2503.03862, 2026a.

Emmy Liu, Kaiser Sun, Millicent Li, Isabelle Lee, Lindia Tjuatja, Jen-tse Huang, and Graham Neubig. What do language models learn and when? The implicit curriculum hypothesis. arXiv:2604.08510, 2026b.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. LLM360: Towards fully transparent open-source LLMs. arXiv:2312.06550, 2023.

Shayne Longpre et al. Consent in crisis: The rapid decline of the AI data commons. arXiv:2407.14933, 2024.

Jonathan B. Losos and Hari Sridhar. Revisiting Losos et al. 1998. Interview, Reflections on Papers Past, 2020. URL https://reflectionsonpaperspast.com/2020/11/11/revisiting-losos-et-al-1998/.

Charles Lovering, Jessica Zosa Forde, George Konidaris, Ellie Pavlick, and Michael L. Littman. Evaluation beyond task performance: Analyzing concepts in AlphaZero in Hex. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Michael Lynch and Jeff Coulter. Discovery and discovering work in the sciences. *Ethnographic Studies*, 21(1):8–32, 2025. doi: 10.26034/lu.ethns.2025.8602.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. DataDecide: How to predict best pretraining data with small experiments. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, 2025.

Fernando Martínez-Plumed and José Hernández-Orallo. Analysing results from AI benchmarks: Key indicators and how to obtain them. arXiv:1811.08186, 2018.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop*, 2020.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in AlphaZero. *Proceedings of the National Academy of Sciences*, 119(47):e2206625119, 2022. doi: 10.1073/pnas.2206625119.

Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Aligning superhuman AI with human behavior: Chess as a model system. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1677–1687, 2020. doi: 10.1145/3394486.3403219.

Robert K. Merton. The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis*, 79(4):606–623, 1988. doi: 10.1086/354848.

James A. Michaelov, Roger P. Levy, and Benjamin K. Bergen. Language model behavioral phases are consistent across architecture, training data, and scale. In *Advances in Neural Information Processing Systems*, 2025.

Shubhra Mishra, Gabriel Poesia, and Noah D. Goodman. From next-token to mathematics: The learning dynamics of mathematical reasoning in language models. In *Second Conference on Language Modeling (COLM)*, 2025.

Akira Murakami. *Individual Variation and the Role of L1 in the L2 Development of English Grammatical Morphemes: Insights From Learner Corpora*. PhD thesis, University of Cambridge, 2014.

Akira Murakami and Theodora Alexopoulou. L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, 38(3): 365–401, 2016. doi: 10.1017/s0272263115000352.

Preetum Nakkiran, Gal Kaplun, Dimitris Kalimeris, Tristan Yang, Benjamin L. Edelman, Fred Zhang, and Boaz Barak. SGD on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems 32*, 2019.

William F. Ogburn and Dorothy Thomas. Are inventions inevitable? A note on social evolution. *Political Science Quarterly*, 37(1):83–98, 1922. doi: 10.2307/2142320.

OpenAI. Working with academic communities to build scientific understanding. Official announcement; ETH section rechecked 9 November 2026, 2026a. URL https://openai.com/index/building-scientific-understanding-with-academic-communities/.

OpenAI. The exponential time hypothesis is false: Scope of the Lean formalization. openai/math, lean/docs/378.md, Scope section, main branch inspected 9 November 2026, 2026b. URL https://github.com/openai/math/blob/main/lean/docs/378.md.

OpenAI. openai/math: README, history, overview and catalog at revision 4eaee87, 2026c. URL https://github.com/openai/math.

OpenAI. On the Navier–Stokes Millennium Prize problem. Official announcement, September 8, 2026, 2026d. URL https://openai.com/index/navier-stokes-solution/.

OpenAI. The exponential time hypothesis is false. OpenAI Math Release preprint, dated 17 October 2026, revised 25 October 2026, 2026e. URL https://github.com/openai/math/tree/main/preprints/The-Exponential-Time-Hypothesis-is-False-October-25-2026.

OpenAI Ireland Ltd. Public summary of training content for GPT-6 Astra, 2026. URL https://cdn.openai.com/pdf/gpt-6-astra-eu-ai-act-public-summary-of-training-content.pdf.

Virginie Orgogozo. Replaying the tape of life in the twenty-first century. *Interface Focus*, 5(6): 20150057, 2015. doi: 10.1098/rsfs.2015.0057.

Gabriele Pallotti and Alessandra Peloso. Acquisition sequences and definition of linguistic categories. *The Open Applied Linguistics Journal*, 1:77–89, 2008. doi: 10.2174/1874913500801010077.

Aðalsteinn Pálsson and Yngvi Björnsson. Unveiling concepts learned by a world-class chess-playing agent. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4864–4872, 2023. doi: 10.24963/ijcai.2023/541.

Kenneth W. Regan and Guy McC. Haworth. Intrinsic chess ratings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 834–839, 2011. doi: 10.1609/aaai.v25i1.7951.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23), 2019. doi: 10.1073/pnas.1820226116.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36*, 2023.

Klaas Sijtsma and Brian W. Junker. A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, 49(1):79–105, 1996. doi: 10.1111/j.2044-8317.1996.tb01076.x. URL https://doi.org/10.1111/j.2044-8317.1996.tb01076.x.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017. doi: 10.1038/nature24270.

Dean Keith Simonton. Independent discovery in science and technology: A closer look at the Poisson distribution. *Social Studies of Science*, 8(4):521–532, 1978. doi: 10.1177/030631277800800405.

Dean Keith Simonton. Multiple discovery: Some Monte Carlo simulations and Gedanken experiments. *Scientometrics*, 9(5–6):269–280, 1986. doi: 10.1007/bf02017248.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. arXiv:2503.19206, 2025.

C. Tristan Stayton. What does convergent evolution mean? The interpretation of convergence and its implications in the search for limits to evolution. *Interface Focus*, 5(6):20150039, 2015. doi: 10.1098/rsfs.2015.0039.

Terence Tao. Mathematics in the age of AI. arXiv:2608.16753, 2026.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 OLMo 2 Furious. arXiv:2501.00656, 2025.

The Stockfish team. Stockfish 14. Stockfish project blog, 2021-07-02, 2021. URL https://stockfishchess.org/blog/2021/stockfish-14/.

Yuandong Tian, Jerry Ma, Qucheng Gong, Shubho Sengupta, Zhuoyuan Chen, James Pinkerton, and C. Lawrence Zitnick. ELF OpenGo: An analysis and open reimplementation of AlphaZero. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, 2019.

Carlos Toledo. Two claimed refutations of the exponential time hypothesis: What was decided, and what was only read. External audit report, summary and sections 1, 5 and 5b inspected 9 November 2026, 2026. URL https://carlostoledo.co/eth/.

Nenad Tomašev, Ulrich Paquet, Demis Hassabis, and Vladimir Kramnik. Assessing game balance with AlphaZero: Exploring alternative rule sets in chess. arXiv:2009.04374, 2020.

George Tsoukalas et al. Advancing mathematics research with AI-driven formal proof search. arXiv:2605.22763, 2026.

Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. PolyPythias: Stability and outliers across fifty language model pre-training runs. In *International Conference on Learning Representations*, 2025.

Erik Y. Wang et al. Measuring progress in reasoning toward mathematical discovery with automatic verification. arXiv:2603.15617, 2026a.

Sophie L. Wang, Amil Dravid, Rulin Shao, Kevin Farhat, Sewon Min, and Alexei A. Efros. Base models can reason by taking a cue from training data. arXiv:2610.06851, 2026b.

Tony T. Wang, Adam Gleave, Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D. Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, and Stuart Russell. Adversarial policies beat superhuman Go AIs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Michael Weisberg and Ryan Muldoon. Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76(2):225–252, 2009. doi: 10.1086/644786.

Robin Whitty. Some comments on multiple discovery in mathematics. *Journal of Humanistic Mathematics*, 7(1):172–188, 2017. doi: 10.5642/jhummath.201701.14. URL https://arxiv.org/abs/1412.3718.

Robert Woods, Dominique Schneider, Cynthia L. Winkworth, Margaret A. Riley, and Richard E. Lenski. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 103(24):9107–9112, 2006. doi: 10.1073/pnas.0602917103.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, 2025.

Tom Zahavy, Vivek Veeriah, Shaobo Hou, Kevin Waugh, Matthew Lai, Edouard Leurent, Nenad Tomašev, Lisa Schut, Demis Hassabis, and Satinder Singh. Diversifying AI: Towards creative chess with AlphaZero. arXiv:2308.09175, 2023.

Siyu Zhang, Yuquan Lu, Yoshua Y. Li, Yifan Ding, Kun Zeng, Yu Huo, Yifu Guo, Rongxiang Weng, Xunliang Cai, and Xiaoying Tang. What RL brings to reasoning: A few decisions the base model would not make. arXiv:2610.24663, 2026.

Kevin J. S. Zollman. The communication structure of epistemic communities. *Philosophy of Science*, 74(5):574–587, 2007. doi: 10.1086/525605.

A The ETH comparison and other candidate doubles

This appendix holds the comparison of the two claimed ETH refutations that motivated the paper (Section 4), with the other candidate doubles we recorded. None of it enters the assessment of the hypothesis.

A.1 The ETH double

OpenAI dates its natural-language ETH disproof to October 16, 2026 and formalization to October 17; Anthropic dates its core algorithm to October 16 (OpenAI, 2026a; Claude and Anthropic, 2026); these are both self-reports of different milestones. Both release commits are dated October 26 in US Pacific time, October 26–27 in UTC (OpenAI, 2026c; Anthropic, 2026b). Neither repository timestamps nor the dates printed on the papers independently establish when their original mathematical content was produced.

For fixed clause width k, the papers state counting algorithms with the same dependence on the number n of variables in the exponential part of their sharp time bounds (OpenAI, 2026e; Claude and Anthropic, 2026):

$$2^{O(n \log\log n/\sqrt{\log n})} \quad \text{and, at fixed clause density,} \quad 2^{O(n/\sqrt{\log n})}.$$

Crucially, matching n-dependence does *not* make the resource or parameter claims identical. OpenAI claims polynomial space and treats k as fixed. Anthropic gives explicit uniform dependence on k and the number m of clauses; its headline theorem does not state a space bound. The formalizations also differ: OpenAI's scope document states a bound with exponential part $2^{O_k(n(\log\log n)^2/\sqrt{\log n})}$, multiplied by a polynomial in the input bit length; the sharper displayed rates remain in an unformalized appendix (OpenAI, 2026b), while Anthropic's formal statement covers its headline time bound using one machine and one constant independent of k, for its specified encoding and input domain: $k \geq 1$, $m \geq 2$, $n \geq 4$, every variable occurring, and no repeated clauses (Anthropic, 2026a). Table 9 separates these scopes from the comparison of the informal arguments.

Anthropic attributes the core algorithm to a run using approximately 14 million output tokens without human input, and claims independent discovery. Its provenance statement also credits researchers working with Claude for improvements to the presentation (Claude and Anthropic, 2026). The OpenAI record gives no ETH-specific compute total. We did not locate the original October 16 discovery artifacts and complete tool traces in the public record checked through November 9, 2026.

An external audit reports successful kernel replays of both released developments and additional checks relating their formal statements to standard machine models (Toledo, 2026).[1] The mathematical comparison below concerns the released arguments; discovery provenance remains a separate question.

A.2 The ladder of sameness

Table 8 compares the target, theorem, quantitative bound, mathematical operation, proof organization, and exposition. For the basic ETH refutation, the papers share classical sparsification and use different engines. Their sharper bounds share a more specific framework: sparsification with a vanishing parameter, a linear number of sparse parity checks, and 2-adic truncation. The recursions nevertheless differ: OpenAI expands into independent smaller components and pays for the number of terms, while Anthropic reduces each term to a smaller connected instance and pays in computation at the leaves.

OpenAI credits Chan–Williams for parity-based constructions and polynomial truncation in deterministic counting. Our comparison uses the 2021 journal version (Chan and Williams, 2021), whose conference precursor appeared in 2016. We did not identify the same sparse-check architecture in the parts examined; its precedents and specificity remain questions for specialist review. Encoding costs in the common representation contribute a $\log\log n$ factor to the upper bounds, without

[1] These are reported external checks; we have not reproduced the builds.

Table 8: The ladder of sameness applied to the two claimed ETH refutations. This compares released arguments, not their discovery histories.

Rung	Agreement	Record
Target	Same	Exponential Time Hypothesis, and #k-SAT counting
Result	Same target conclusion	An ETH refutation, each accompanied by a Lean development with a different scope
Time bound	Same rate in n at fixed k	Exponents $n \log\log n/\sqrt{\log n}$ in general; $n/\sqrt{\log n}$ at fixed density
Other guarantees	Different	OpenAI: polynomial space at fixed k. Anthropic: uniform explicit dependence on k, with no headline space bound
First move	Shared	Count-preserving sparsification at a constant parameter, from the same sources; generic
Framework (sharp bound)	Shared	Sparsify with $\varepsilon \to 0$; $O(n)$ sparse parity checks selected by the paper's counter; 2-adic truncation; contract W wires at cost $2^{O(W/\sqrt{\log W})}$. Some ingredients come from Chan–Williams; we did not identify this sparse architecture in the parts examined
$\log\log n$ step	Shared encoding cost	Both name $O(n)$ marked incidences in $\lceil \log \binom{N}{d} \rceil$ bits. This comparison of upper bounds establishes no necessity result
Engine for $\sqrt{\log}$	Different	OpenAI: 64 independent children, paying in their number. Anthropic: one smaller instance, paying at brute-force leaves
Main-body route	Different	OpenAI: interpolation of replacement forms. Anthropic: a low-rank identity for crossings, applied recursively
Few-violation regime	Different	Tanner codes against truncated inclusion–exclusion
Exposition	Citation direction	Anthropic's paper cites OpenAI's preprint; the inspected OpenAI version does not cite Anthropic. This dates an opportunity for revision, not a discovery route

establishing a lower bound or the factor's necessity. Its appearance therefore adds little independent evidence of a common discovery route. Shared literature, mathematical structure, and later revisions could each contribute to the resemblance, whose description depends on the level of comparison (Orgogozo, 2015).

A.3 Twelve source checks

Table 9 summarizes twelve checks against the ETH papers (OpenAI, 2026e; Claude and Anthropic, 2026). Page and section locators and the algebraic checks are in the companion repository.

OpenAI's main-body interpolation and sharper tensor-contraction appendix use different constructions, as do Anthropic's main proof and per-bit sketch. The swap connection joins OpenAI's main body to Anthropic's sketch. At the sharp bound, we identify related weighted counting structures.

The common rates refer to fixed clause width and the stated density regimes; parameter dependence, space claims, and formalization scope differ.

Table 9: Twelve source checks in the ETH comparison. This is not a reproduction of the formal developments.

ID	Finding
C1	OpenAI's main body uses rational interpolation; its sharper bound uses tensor contraction in Appendix D.
C2	OpenAI's Lean scope has an additional $\log \log n$ factor relative to its sharp informal bound, with polynomial space at fixed k. Anthropic states its headline time bound uniformly in k on an explicit encoded input domain, without a headline space bound.
C3	OpenAI's main-body construction organizes exchanges by their common eigenspaces and an interpolation window.
C4	Anthropic's main body uses an eight-square identity on three-bit blocks and a global merge choice.
C5	Anthropic's per-bit sketch uses a swap identity. Its rank count differs from the simpler count carried through the main proof.
C6	The swap connection is between OpenAI's main body and Anthropic's accompanying sketch, not between both main proofs.
C7	The sharp arguments share count-preserving sparsification, sparse parity checks, greedy selection, and a global 2-adic truncation budget, with different constructions and load control.
C8	The sharp square-root savings involve related weighted counting structures; their mathematical roles differ.
C9	Twenty-eight cited works were matched across the bibliographies, allowing conference and journal versions; many construction-specific precedents differ.
C10	Neither paper cites Alman–Vassilevska Williams.
C11	The main bodies prove different levels of result; the appendices are essential to comparing the headline theorems.
C12	Public release dates and self-reported milestones describe different events. Original discovery traces were not located in the public record checked through November 9, 2026.

A.4 Dated resolutions across labs

We also surveyed dated resolutions across labs: the 721 manuscripts, in 373 families, of OpenAI's release at the recorded revision, curated events, and related results from other labs.

Table 10 records three candidate machine–machine cases and one human–machine case. In this survey, ETH is the only candidate for which primary mathematical texts were inspected on both

machine sides. The Erdős cases have less complete records; one later announcement follows a public result with the run date unstated and a different quantitative bound. The human–machine random k-SAT case has acknowledged common antecedents.

Table 10: Candidate cases of overlapping discovery in the survey of dated resolutions. Dates mix reported milestones, announcements, and manuscript dates as marked; independence and equivalence of results require separate evidence.

Problem	Parties	Reported dates	Standing
ETH	OpenAI; Anthropic	Both report milestones on 2026-10-16	Primary text inspected on both sides; the matching dates describe different events and are self-reported
Erdős #846	DeepMind prover; OpenAI internal	2026-02-21 to 25	A curated wiki states independence; original discovery records were not verified
Erdős #90	OpenAI internal; Claude Mythos	Announced 2026-05-20; 05-26	Later claim follows the public one by six days, with a weaker exponent and unstated run date; independence is not established
Random k-SAT threshold	Carenini; OpenAI	2026-10-05; manuscript 09-25, released 10-06	Human–machine; OpenAI acknowledges priority. Common antecedents offer a possible explanation of proximity

The 3SUM/APSP paper credits Claude with the core algorithm and its human authors with checking, simplifying, strengthening, and extending the result (Alman and Vassilevska Williams, 2026). Two of Anthropic's reported results also arose during work on other tasks, showing that a machine discovery need not begin with a prompt naming its eventual target. Estimating the frequency of multiples would additionally require a defined problem population, systematic checks for other discoverers, unsuccessful attempts, and a fixed time window.

A.5 Navier–Stokes as a confounded case

OpenAI's primary Navier–Stokes announcement describes an effort beginning on September 1 after rumors about solved Millennium problems, reaching a claimed proof after about 88 hours, and producing approximately 130 billion output tokens in the Navier–Stokes agent trajectories (OpenAI, 2026d). These are the lab's reported process and resource figures. The account also describes human consolidation of agent insights and follow-up prompts. Buckmaster describes how information about ongoing work circulated (Buckmaster, 2026).

The signal may explain the timing of the attempt; the amount of mathematical information transferred is unknown. Together with the large differences in reported compute and human involvement, this makes announcement order a poor measure of accessibility under fixed conditions.

B Prespecified plans and post hoc additions

File paths in this appendix and the next are relative to the companion repository.

The Leela plan was fixed at commit `a91384f`, before item-level comparisons. It sets the pairs, endpoints, gates, and the anchor/test split. Dated amendments in `docs/protocols/leela-analysis-plan.md` record later changes to run definitions, numerical conventions, exclusions and method contrasts, each stating what had been observed.

The order gates require a reversal upper bound of at most .10 and a coverage lower bound of at least .20. A strength-matched comparison is graded only when the common range spans at least .20 of anchor solve share and holds at least four distinct strength values per curve; otherwise it is descriptive only.

The plan's four questions concern order, incremental prediction, budget, and overlap at matched strength. Incremental prediction is the reduction in held-out log loss when another run's accessibility is added to item features and difficulty proxies. Bank-specific strength axes are isotonic fits to 201 Lichess and 56 tablebase anchors, excluded from tests.

The open-problem plan, `docs/protocols/open-problems-overlap-plan.md`, was committed at `09e7d70`, before per-conjecture outcomes were opened. It fixed overlap, ladder ordering, scaffold comparisons, and comparisons within and across labs, using upstream results pinned to `fd09021e`. The run at `38d2789` reproduced the published counts before reporting the planned analyses.

The readings of the presentation follow-up were committed at `fc1bd1c`, after the format split was known and before follow-up outcomes were examined. An amendment at `4147ded` followed inspection of anchors showing lost strength range under blank history planes; it kept the thresholds, so that undefined accessibility is reported as a partial result.

The engine was Lc0 v0.32.1, source commit `fd71a2d9`. Hashes for the binary, network catalog, derived axis, and 528 raw record files are recorded with the analysis; the raw manifest is `data/chess/results/leela/raw-manifest.sha256`. The Stockfish, compute-invariance, evaluator-toggle, and Leela run protocols are in `docs/protocols/`. The ETH hint-length protocol is a proposed study and was not run.

B.1 Post hoc additions

The following were added after the planned outcomes were known and carry no confirmatory status. On open problems: the checks on jointly solved pairs and flagged statements, the one-scale model, and the acquisition decomposition (`scripts/open_problems/acquisition_decomposition.py`). On chess: the 16-visit own-range and format comparisons and the acquisition decomposition (`scripts/chess/lc0/acquisition_decomposition.py`). The decompositions' cohort restrictions, pair strata, and uncertainties are descriptive; they do not amend the prespecified populations or decisions.

Propositions 1 and 2, Table 1, the summary by claim in Table 7, and the dated forecast in Section 8.3 were likewise written after every result reported here was known. The pair counts for the four full-set models, and the attribution of 14 of 22 Guttman errors to one pair, are read from the saved solve patterns of the scale check.

A candidate protocol for the direct test of Section 8.1 is at `docs/protocols/prospective-acquisition-v1-2026-11-09.md`. It has not been adopted, executed or preregistered; a final protocol awaits selection of suitable trajectories and an assessment of feasibility and power.

C Chess design and implementation

C.1 One lineage: Stockfish

The Stockfish pilot evaluated 39 configurations on both banks in 33.3 DTU core-hours, testing the instrument and tracing one lineage. Design and results are in `docs/protocols/chess-pilot.md` and `notes/chess-pilot-results.md`.

Across eight configurations, median Spearman agreement between the original and color-flipped Lichess item orders was .89, with an observed range of .81–.95, using depth-grid first-move thresholds.

Table 11: Compute invariance on Stockfish: excess reversal rate, with 95% interval.

Bank	Node budgets	Depth budgets
Tablebase	0.0096 (0.0058, 0.0147)	0.0285 (0.0164, 0.0432)
Lichess	0.0077 (0.0045, 0.0121)	0.0474 (0.0298, 0.0670)

This range describes variation between configurations, not a confidence interval. Against that reference, order drifts across versions from 2013 to 2026. Comparisons around the introduction of Leela-derived data (The Stockfish team, 2021) show modest observed differences, but do not identify or bound its causal effect. With search fixed inside a binary, evaluator changes move the order at Stockfish 16, and the network adds a step at Stockfish 12.

Compute produces small positive excess reversals, particularly under node budgets (Table 11). Older versions enforced low node limits differently; exact-limit builds corrected the affected results and motivated budget validation for Leela.

C.2 Separate historical runs: Leela

The main design sampled 24 checkpoints from each of eleven historical training runs, giving 264 networks. T76 and T77 were designated as the primary pair, with nine further runs compared against each reference. The first-move procedure used one visit for every network and 16 visits for the primary pair. The full run comprised 799,344 searches and used 16.2 core-hours. Table 12 records the runs and their relevant implementation differences.

T77's first training window is undocumented; independent ancestry and disjoint self-play data are unestablished.

Anchor validation checked exact visits, determinism, history, and the one-visit policy endpoint across three CPU models. Disabling the cache changed four of 4,112 moves at 16 visits, so the recorded endpoint uses classic Lc0 search with its cache enabled and cleared between items.

Ten forced-move tablebase test items were excluded before comparison. A further item, with two knight checks and no legal predecessor, produced a disagreement between Lc0's move list and the rules and was excluded everywhere. Eight other unreachable positions were retained. The final test banks contain 493 tablebase and 1,801 Lichess items.

T76's classical format uses eight plies of history; T77's canonical format stops at the last irreversible move and applies a symmetry transform. Bare positions give the former repeated boards and the latter potentially one board. Altering history changed one to 17 of 201 anchor first moves in validation.

Strength matching uses anchors alone. For each bank, an isotonic fit maps the sampled checkpoints to anchor solve share. Item accessibility integrates solved indicators over the common strength range. A resolution rule excludes comparisons in which too few checkpoints span that range. On tablebase anchors at one visit, T76 passes T77's ceiling within four consecutive networks, so the primary pair fails that rule. The prespecified 16-visit comparison, whose range spans all four curves, also fails it on the tablebase bank. The later comparison of the two 16-visit curves uses its own range and is post hoc. Lichess supports the intended strength comparisons.

Chronological acquisition follows training order through a fixed endpoint, with censoring represented as intervals, independently of the anchor strength axis.

The main bounds use 2,000 paired resamples of test items, preserving rating-band strata on Lichess and material-by-DTZ strata on the tablebase bank. Complete item trajectories stay together across all nets and runs; the millions of derived pairs are not independent sampling units. Every

Table 12: Leela runs in the study, with the solve counts of each run's last selected net on anchor items, already seen in the range check. Anchors never enter a test.

Run	Role	Size	Input format	Lichess (of 201)	Tablebase (of 56)
T76	primary	15×192	classical	149	45
T77	primary	15×192	canonical	150	31
T75	replication	15×192	classical	148	43
T79	replication	15×192	classical; attention head, mish	146	43
T40	replication	20×256	classical	137	45
T71	replication	19×256	canonical	142	30
T74	replication	10×128	classical	129	41
T72	replication	10×128	canonical	130	27
T70	replication	10×128	classical	121	38
T59	replication	10×128	classical	136	41
T50	replication	10×128	classical	117	40

Lichess test item has a distinct game identifier in this bank. The 200 anchor resamples separately refit the strength axes and ranges. They are a sensitivity analysis, not joint item-and-anchor confidence intervals. Predictive bounds resample stored out-of-fold item losses and condition on fitted predictors, penalty selection and checkpoint selection. No resampling step samples new training lineages.

The prespecified anchor safeguard is asymmetric: it can downgrade a supported accessibility comparison to inconclusive, while leaving a contradicted status conditional on the original axis. For the same-format T40–T76 comparison, the separate anchor-sensitivity lower reversal endpoint is .113; for T50–T76, 16 of 200 anchor resamples fail axis resolution. Their original accessibility statuses are preserved. Their chronological comparisons, which do not use an anchor axis, are both inconclusive.

Post hoc, final canonical networks solved 27–31 of 56 tablebase anchors and classical networks 38–45, with format confounded by era and recipe. The presentation follow-up and its partial tablebase result are reported in Section 6.3.

C.3 Detailed results

Table 13 collects the accessibility, chronological, predictive, and sensitivity results summarized in Section 6.3.

C.4 Acquisition within the observed window: a post hoc decomposition

The decomposition reconstructs the primary T76–T77 comparisons from the saved records, reproducing all four accessibility and chronological pair counts on both banks, and classifies each item in each run as left-censored, observed inside the window, or right-censored. The interior subset requires an observed acquisition in both runs: at least one sampled failure followed by a nonempty final uninterrupted solve suffix. An item first meeting this condition at the final checkpoint remains eligible. Pair order uses the original strict interval-separation convention, including its exclusion of adjacent intervals with a common boundary.

Table 14 partitions the complete chronological comparison into pairs whose two items are in this interior subset and all remaining pairs. The latter involve at least one censoring classification in at least one run; they still provide valid order information. The interior populations contain 1,105 Lichess and 212 tablebase items. On Lichess, their .118 reversal rate has exploratory item-bootstrap

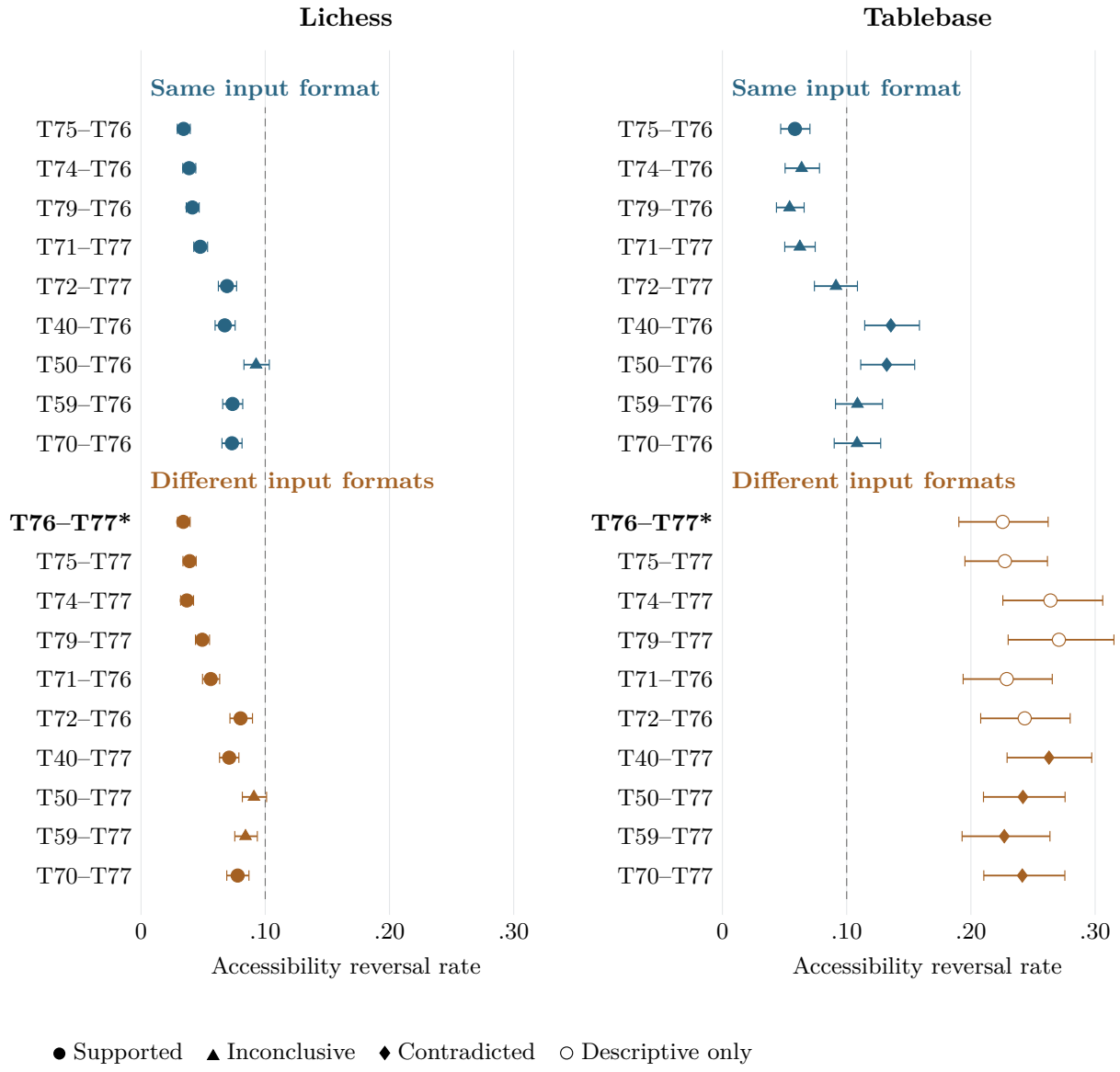


Figure 3: Accessibility reversals at one visit (item-order margin .10), with one-sided 97.5% bounds and the .10 reversal threshold. The primary pair (*) and all 18 replication comparisons are grouped post hoc by input format. Symbols preserve the recorded decisions, including anchor-sensitive inconclusive results; open circles fail the axis-resolution rule and are descriptive only. Coverage lower bounds exceed .20 throughout. Chronological results appear in Table 13.

bounds of approximately .102–.135; on tablebase, the corresponding rate is .172 with bounds .123–.224. Coverage within the selected populations is .720 and .634, respectively; the table instead uses all original pairs as the common denominator.

The subgroup was defined after reading the full results. Selection depends on the same outcomes whose order is compared, so the bootstrap describes a fixed selected subgroup conditional on the recorded histories; it does not adjust for selection or support population-wide inference about acquisition interiors. No original decision status is replaced. Both the original censored analysis and this diagnostic are reported because they answer different descriptive questions.

The reconstruction, source hashes, pair-level censoring strata and item lists are in `scripts/chess/lc0/acquisition_decomposition.py` and `data/chess/results/leela/analysis/review-2026-11-09/`. The raw records, plans and primary results are unchanged.

Table 13: The Leela analysis, with margin 0.10 for accessibility. Cells give reversal (one-sided 97.5% bounds), then coverage, or Δ in bits with 95% intervals. Bounds concern item sampling conditional on recorded histories; predictive bounds also condition on fitted predictors. Tablebase at 1 visit is descriptive only for T76–T77 by the resolution rule.

	Lichess (1,801 items)	Tablebase (493 items)
Accessibility, T76–T77, 1 visit	supported: .034 (.029, .040), .730	descriptive only: .226 (.190, .262), .559
Chronological order	supported: .081 (.071, .090), .749	contradicted: .236 (.201, .272), .624
Incremental prediction Δ, each direction	.310 (.289, .332); .322 (.301, .344); substantial	none graded on the bank axis; on the Lichess axis .092 (.062, .122), .068 (.043, .096)
16 visits, prespecified range	supported: .056 (.048, .063), .695	descriptive only: .221 (.187, .257), .576
16 visits, own common range (post hoc)	supported: .047 (.041, .054), .723	contradicted: .191 (.161, .221), .639
1 against 16 visits within a run	T76 .079, T77 .081; supported	descriptive only: T76 .126, T77 .118
Matched strength, new solves against expected	272/124, 180/81, 198/129, 113/92; all above the null band	unavailable: inadequate range
Accessibility replications, 18 dependent comparisons	15 supported, 3 inconclusive, 0 contradicted	1 supported, 6 inconclusive, 6 contradicted, 5 descriptive only
Same format, 9 pairs, reversal	.034 to .093	.054 to .136
Different format, 10 pairs, reversal	.034 to .091	.226 to .271
Same-format T40–T76 and T50–T76, chronological	—	.106 (.087, .127), .686; .101 (.081, .122), .637; both inconclusive
Presentation check, format gap (post hoc follow-up, readings fixed first)	bare: .005, .006 on 697 matching items; no split	blank planes: undefined, .176; partial
Method contrast, input format	−.005, .000, −.008	withheld
Method contrast, policy head and activation	−.007 in reversal, +.017 in coverage	+.004 in reversal, +.026 in coverage

Table 14: Post hoc decomposition of the primary chronological comparison at one visit. "Interior" requires both items' acquisitions to be observed inside both runs' windows. "Any censoring" is the complementary pair population. Coverage always uses all original pairs: 1,620,900 on Lichess and 121,278 on tablebase. The interior subsets are selected by observed outcomes; no primary decision status is replaced.

Bank	Pair population	Concordant	Reversed	Reversal	Coverage
Lichess	All pairs	1,116,096	97,754	.081	.749
	Interior	386,974	52,002	.118	.271
	Any censoring	729,122	45,752	.059	.478
Tablebase	All pairs	57,774	17,856	.236	.624
	Interior	11,745	2,438	.172	.117
	Any censoring	46,029	15,418	.251	.507