RESEARCHTREE: MEASURING THE RESEARCH HORIZON OF LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

How capable are language models at research? We measure one part of this question: how much of existing human research a model can reconstruct on its own, from a single step up to a whole paper. We introduce RESEARCHTREE, which turns a research paper into a tree of self-contained problems, from its main result down to atomic steps with solutions taken from the paper. The model under test answers every problem independently, and a judge scores each answer against the paper's derivation beneath it. Two measurements follow. The *research horizon* is the problem size, in atomic steps, at which a model's expected score falls to one half. The *decomposition curve* shows how much of a paper a model recovers as the paper is given to it in smaller and smaller parts, and its *deficit* is the share the model still misses on average. On 20 recent theoretical papers and ten models, frontier models solve almost every atomic step but much less of each whole paper. Research horizons range from under one step to about 40, differ several-fold between fields, and across four leading releases have doubled roughly every eight months, a trend that rests heavily on the earliest release, while the decomposition deficit has halved about once a year. Posed a whole paper, models mostly leave parts of the derivation out rather than get them wrong. Because the trees are built automatically from new papers, the evaluation can be renewed after every model release and could extend to other fields. We release RESEARCHTREE as a public dataset: 20 paper trees with 1,079 problems (744 atomic steps with reference solutions from the paper), 10,790 model answers and 32,880 graded judgments, together with the code to build new trees.

1 INTRODUCTION

Language models are now used to prove lemmas, derive formulas and draft whole papers (Lu et al., 2024; Gottweis et al., 2025; Mitchener et al., 2025). The central question is how capable they are at research. Research is hard to evaluate directly, but its results are written down: a paper states a main result and records the derivation that establishes it. We therefore ask a narrower question that can be measured: how much of an existing piece of human research can a model reconstruct on its own, and how does this change as the piece grows from a single step to a whole paper? For software agents, Kwa et al. (2025) answer a similar question with a *time horizon*, the length of task, in human working time, that an agent completes half the time. Research has no such measure. Research benchmarks score each task at one or two fixed granularities, whether a paper to replicate (Starace et al., 2025), a problem with its sub-steps (Tian et al., 2024), or a research question, proof or exam item (Miao et al., 2026; Glazer et al., 2024; Phan et al., 2025). A single score cannot tell a model that proves every step of a derivation but cannot put them together from a model that proves nothing, and it cannot say how much more the first model would recover if the paper were split up for it. These benchmarks are also written by experts, one problem at a time, so they are expensive to extend, and they saturate or leak into training data within a few model generations (Ott et al., 2022; Xu et al., 2024).

Our observation is that a paper already carries both a measure of size and a grading key. Its main result rests on intermediate results, which rest on smaller claims, down to steps a graduate student could do as one exam question, and the paper writes out the derivation of each. RESEARCHTREE makes this structure explicit (Figure 2). A builder model turns a paper into a tree of self-contained problems, with the main result at the root and atomic steps, whose solutions come from the paper, at

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the leaves. The model under test answers every problem on its own, and a judge scores each answer by how much of the paper's derivation beneath it the answer covers. Nothing in this recipe is specific to one field: any paper that writes out its derivation can in principle be turned into a graded tree in the same way, and we test it here on theoretical papers. Because new papers appear every day, a fresh set can be built after any model's training cutoff, much as live benchmarks refresh their items (White et al., 2025; Jain et al., 2025a).

The scored tree gives two measurements. The first asks how large a part of a paper a model can reconstruct alone, relating score to problem size, the number of atomic steps a problem contains. Scores fall along a logistic curve in log size, and the size at which a model's expected score falls to one half is its *research horizon* (Figure 1), an analogue of the time horizon that needs no human timing. The second asks how much of a whole paper a model recovers when the paper is given to it in parts, from the whole paper at once to all of its steps one by one. This *decomposition curve* shows how far a model is from reconstructing the paper and how much of that distance the paper's own structure closes; its *decomposition deficit*, the share of the paper a model still misses averaged over all ways of splitting it, summarises it in one number.

We apply ResearchTree to 20 theoretical papers from September 2026 in combinatorics, computational complexity, statistical mechanics and high-energy theory, and to ten models from two labs; theoretical papers are cheap to evaluate, since every step is a derivation that needs no experiments or compute beyond a Python sandbox. The strongest models solve almost every atomic step, but even the best model averages only about 0.6 when asked for a whole paper. Research horizons range from under one step to about 40 steps, and the same model reaches problems several times larger in physics than in combinatorics or complexity. Once problem size is measured in units of each model's horizon, all capable models follow approximately the same curve. Across four leading releases, the horizon has doubled roughly every eight months, though this rests heavily on the earliest release, and the decomposition deficit has halved about once a year. When a model is posed a whole paper, the judge mostly finds parts of the derivation left out rather than errors.

Unlike research benchmarks that pose problems at one or two fixed granularities (Starace et al., 2025; Tian et al., 2024; Glazer et al., 2024; Zhang et al., 2025a), and time horizons that need timed

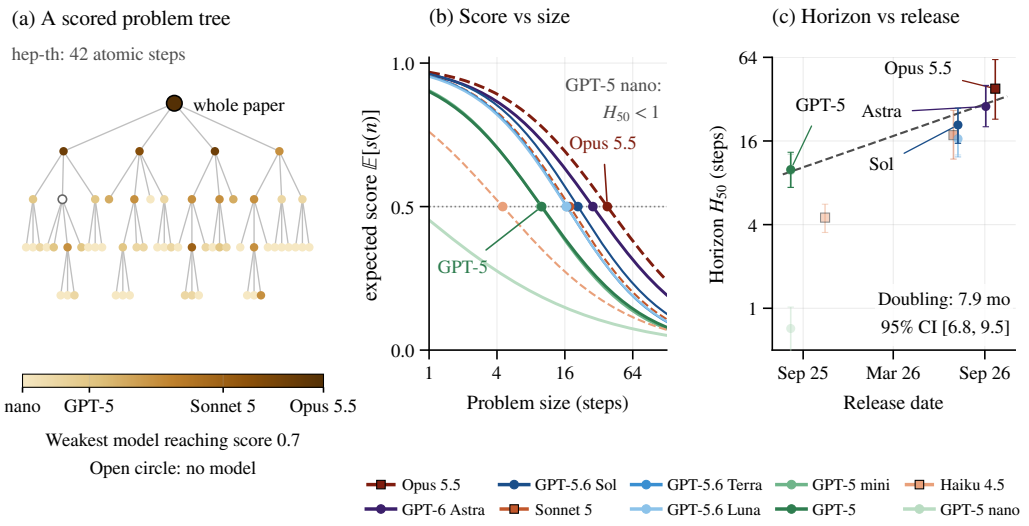

Figure 1: **The research horizon.** (a) A real tree built from a 42-step high-energy theory paper. Each node is coloured by the weakest model that scores at least 0.7 on it; weak models succeed only near the leaves, and only Claude Opus 5.5 reaches 0.7 on the whole paper (0.70, against 0.69 for GPT-6 Astra). (b) Across all 20 papers, a model's expected score falls with problem size, counted in atomic steps; the size where the fitted curve crosses one half is its research horizon H_{50} (dots). (c) Research horizon against release date, with 95% intervals from resampling papers (squares: Claude; circles: OpenAI). The dashed line fits the four leading releases.

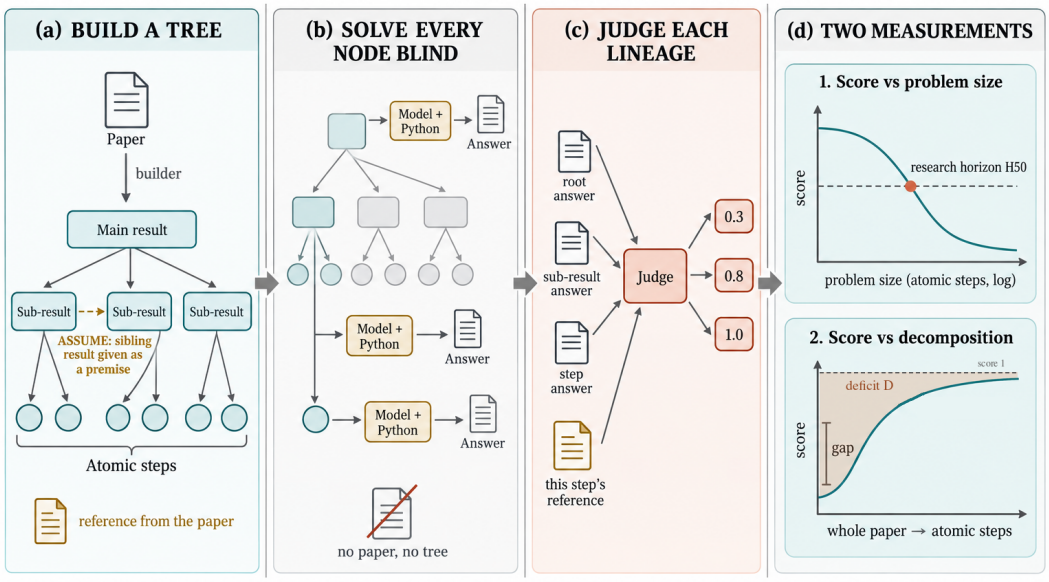

Figure 2: **Overview of RESEARCHTREE.** (a) A builder model turns a paper into a tree: the main result at the root, weighted sub-results inside, and atomic steps with solutions taken from the paper at the leaves. A result that one sub-problem needs from a sibling is given to it as an assumption. (b) The model under test answers every node from that node's statement alone. (c) For each leaf, a judge scores every answer on the path from the root to that leaf against the leaf's reference solution. (d) The scored tree gives two measurements: score against problem size, whose one-half crossing is the research horizon, and score against how finely the paper is decomposed, whose shortfall averaged over the curve (shaded) is the decomposition deficit D. The tree, scores and curves in this figure are illustrative, not data.

human baselines (Kwa et al., 2025), RESEARCHTREE poses every level of a paper as its own problem and measures size in the paper's steps (related work: Appendix A).

Contributions.

- We introduce RESEARCHTREE, an automatic pipeline that turns a research paper with a written derivation into a graded evaluation: a tree of self-contained problems, from the main result down to atomic steps, scored against the paper's own derivation. Since new papers appear every day, the evaluation can be renewed with fresh problems.
- We propose two measurements on the scored tree: the research horizon and the decomposition curve with its deficit, both free of success thresholds.
- We evaluate ten models on 20 papers: both measurements follow curves shared across models, differ strongly between fields and improve across leading releases.
- We release RESEARCHTREE as a public dataset (20 trees, 1,079 problems, 10,790 answers, 32,880 judgments) with the code to build new trees; it could also serve as a graded environment for training models on long-horizon research.

2 RESEARCHTREE

RESEARCHTREE builds a tree of problems from a paper, has the model under test answer every problem independently, scores each answer against the paper's own steps, and reads two measurements off the scored tree (Figure 2).

2.1 TREES OF SELF-CONTAINED PROBLEMS

Root and decomposition. A builder model (Claude Opus 5.5; Claude Opus 5 for two earlier trees, Section 3) reads the full paper and states its main result as one problem, with every symbol defined and no hint of the method. It then splits each problem into two to four sub-problems along the logical dependencies of the argument and gives each an importance weight, with weights summing to one among siblings. Sub-problems ask for results; they do not reveal them.

Self-containment. The model under test sees one problem and nothing else, so every problem must stand alone: the builder is told to define every symbol and never to refer to the paper, the parent or a sibling. When a sub-problem needs a result proved in a sibling, it states that result as an assumption ("Assume that $E(b)$ satisfies …; find $E(b)$"), as in the tree of Appendix D (Figure 8). This keeps problems independent, but it means a sub-problem can receive for free a result that its parent must derive; the difference between scores at the two levels therefore also reflects what decomposition hands the model.

Atomic steps and reference solutions. Decomposition stops when a problem is atomic, meaning a well-prepared graduate student could answer it as one exam question; we call these leaves the *atomic steps* of the paper and measure size in them. For each leaf, the builder writes a reference solution from the paper, filling in steps the paper leaves implicit and marking which steps it inferred. Only leaves have reference solutions. They define what the judge looks for, so an answer that reaches a result by a different valid route can receive little credit (Section 5).

2.2 BLIND EVALUATION

The model under test answers each problem from its statement alone, with a Python sandbox and without the paper or the tree, and is asked for a complete derivation. Each problem is answered once, and each answer is judged against every leaf beneath its node. For each leaf L with reference solution y_L and path (*lineage*) $n_0, \dots, n_d = L$ from the root, a judge scores each answer on the path against y_L,

$$\mathrm{Score}(n_k, L) = J\big(a(n_k),\ \mathrm{problem}(L),\ y_L\big) \in [0,1], \qquad k = 0, \dots, d. \tag{1}$$

The judge asks how much of what step L requires is derived in the answer to the larger problem n_k. The judge, Claude Sonnet 5, uses a five-band rubric from "unrelated or contradicting" (0–0.1) to "all key elements" (0.9–1). A node with m leaves beneath it thus receives m scores, each against a concrete reference.

2.3 SCORES AND MEASUREMENTS

Node score. Let $\mathcal{L}(n)$ be the leaves beneath node n and $w(n \to L)$ the product of importance weights on the path from n to L. The score of a node is the weighted average of its scores against its leaves,

$$s(n) = \frac{\sum_{L \in \mathcal{L}(n)} w(n \to L)\, \mathrm{Score}(n, L)}{\sum_{L \in \mathcal{L}(n)} w(n \to L)}, \qquad s(L) = \mathrm{Score}(L, L). \tag{2}$$

The *whole-paper score* $R_{\mathrm{root}} = s(n_0)$ is the root's score, and the *step score* R_{full} is the same weighted average over the leaves' scores on their own problems; we call $R_{\mathrm{full}} - R_{\mathrm{root}}$ the *composition gap*.

Measurement 1: score against problem size. The size of a problem is the number of atomic steps beneath it, $|\mathcal{L}(n)|$. For each model we fit a logistic curve of the node score in log size over all nodes of all papers, each node weighted equally,

$$\mathbb{E}\big[s(n)\big] = \sigma\big(\alpha + \beta \log_2 |\mathcal{L}(n)|\big), \qquad H_{50} = 2^{-\alpha/\beta}, \tag{3}$$

directly on the scores, without a success threshold. The *research horizon* H_{50} is the size at which the fitted expected score is one half. Fitting the same curve within one field gives a field-specific horizon. For comparison with time horizons we also fit the probability that a score reaches 0.7; it ranks models almost the same way, with one adjacent swap (Appendix H). All intervals resample papers.

Measurement 2: score against decomposition. A *cut* F is a set of nodes that together cover every leaf exactly once; scoring it corresponds to posing the paper as the problems in F,

$$R(F) = \sum_{n \in F} w(n_0 \to n)\, s(n). \tag{4}$$

The cut of all leaves gives R_{full} and the root alone gives R_{root}. Starting from the leaves, we repeatedly merge a complete set of siblings into their parent, deepest first, until only the root remains. Plotting $R(F)$ against the fraction of the tree still decomposed, $x = (|F| - 1)/(N - 1)$ for a tree with N atomic steps, gives the *decomposition curve*; we interpolate each curve linearly in x and average ten random merge orders (details and pseudocode in Appendix E). We summarise it by the *decomposition deficit* $D = 1 - \int_0^1 R(F_x)\, dx$, the score a model still misses on average over all ways of posing the paper, from the whole paper to its steps; like the horizon, it needs no success threshold. We also report how much of the composition gap each level of the tree recovers.

3 Experimental setup

Papers and trees. We start with theoretical papers because they are the cheapest to evaluate at scale: every step is a derivation that needs only a Python sandbox, not datasets, GPUs or a laboratory; an empirical paper decomposes in the same way but needs an execution environment for its leaves. We use 20 theoretical arXiv papers posted in September 2026, five each from combinatorics (`math.CO`), computational complexity (`cs.CC`), statistical mechanics (`cond-mat.stat-mech`) and high-energy theory (`hep-th`), chosen by an LLM filter for results that can be checked by derivation alone (Appendix G). They appeared after the release of nine of the ten models; we did not audit training data. The trees have 1,079 problems in total, of which 744 are atomic steps; papers have 15 to 130 steps and trees are three to six levels deep (Appendix G). The builder decides itself when a problem is atomic; only at the maximum depth of six, which four trees reach, is it required to stop. Most reference solutions fill in at least one step the paper leaves implicit (87% carry an inferred-step marker), and 21 are the builder's reconstructions of results the paper does not spell out; no domain expert audited them. Two trees come from an earlier builder run (Appendix G).

Models. We evaluate ten models from two labs, released over 13.5 months and spanning three tiers of each lab's line-up: GPT-5, GPT-5 mini and GPT-5 nano; GPT-5.6 Sol, Terra and Luna; GPT-6 Astra; and Claude Opus 5.5, Sonnet 5 and Haiku 4.5 (release dates in Table 1). Each runs in its own lab's agent harness, Claude models in Claude Code and OpenAI models in Codex, with the same task prompt and a Python sandbox without network access; OpenAI models use high reasoning effort. The model is instructed, not technically prevented, to read only its own problem. Every model answers all 1,079 problems, so all comparisons are paired.

Judge. Claude Sonnet 5 judges all answers; it may consult the paper for notation only. Because it shares a family with three of the models, GPT-5.6 Sol re-judges four papers as a check (Appendix C).

Failures, provenance and cost. Truncated or corrupted answers received one retry and remaining failures count as answers; serving metadata matched every requested model, and solving cost $0.01 to $1.27 per problem (Appendices M, N and O).

4 Results

4.1 Whole papers are much harder than their steps

Posed one atomic step at a time, the papers are almost solved: eight of the ten models score above 0.9 on the steps (Table 1). Posed as a single problem, they are not: even the best model, Claude Opus 5.5, averages 0.62 on the whole papers, and most models stay below 0.5. The worked example in Appendix D, a paper on the Komlós conjecture, shows the pattern on one paper: asked for the whole theorem, Opus 5.5 and GPT-6 Astra prove its two easier parts and leave out almost all of the main proof, although they solve every step of that proof when the steps are posed one by one.

Table 1: **Summary by model.** Whole-paper score R_{root} and step score R_{full} (means over 20 papers); research horizon H_{50} (steps) and its variant $H_{50}^{\tau=0.7}$ based on a 0.7 success threshold; decomposition deficit D (mean over papers); difference in whole-paper score from Sonnet 5; list-price cost per problem. Brackets are 95% intervals from resampling papers (simultaneous intervals: Table 11); differences are taken from Sonnet 5, the judge model; draws: 1,000 for H_{50}, 2,000 for $H_{50}^{\tau=0.7}$ and D, 10,000 for the differences.

Model	Released	R_{root}	R_{full}	H_{50} [95% CI]	$H_{50}^{\tau=0.7}$ [95% CI]	D [95% CI]	Δ vs Sonnet 5 [95% CI]	$/node
Opus 5.5	2026-09	0.62	0.98	38.1 [23.0, 61.7]	23.7 [13.2, 44.2]	0.11 [0.09, 0.13]	+0.15 [+0.10, +0.20]	0.77
GPT-6 Astra	2026-09	0.51	0.97	28.3 [20.3, 40.2]	17.9 [11.0, 30.2]	0.13 [0.11, 0.15]	+0.04 [-0.06, +0.12]	0.51
GPT-5.6 Sol	2026-07	0.49	0.97	20.9 [15.4, 27.6]	12.0 [8.7, 18.1]	0.14 [0.12, 0.16]	+0.02 [-0.01, +0.06]	0.29
Sonnet 5	2026-06	0.47	0.97	17.6 [11.9, 26.6]	9.3 [6.6, 14.1]	0.16 [0.13, 0.19]	reference	1.27
GPT-5.6 Luna	2026-07	0.44	0.96	16.5 [12.7, 22.1]	9.1 [6.8, 13.0]	0.16 [0.14, 0.19]	-0.03 [-0.09, +0.02]	0.09
GPT-5.6 Terra	2026-07	0.44	0.97	16.6 [12.3, 21.5]	9.4 [7.3, 12.3]	0.16 [0.14, 0.19]	-0.03 [-0.10, +0.02]	0.15
GPT-5 mini	2025-08	0.35	0.92	9.8 [7.5, 12.9]	4.3 [3.2, 5.8]	0.24 [0.20, 0.28]	-0.12 [-0.16, -0.08]	0.05
GPT-5	2025-08	0.35	0.92	10.0 [7.4, 13.3]	4.4 [3.5, 5.8]	0.25 [0.21, 0.29]	-0.12 [-0.21, -0.06]	0.25
Haiku 4.5	2025-10	0.26	0.77	4.5 [3.5, 5.6]	1.5 [1.2, 1.7]	0.41 [0.37, 0.45]	-0.21 [-0.27, -0.16]	0.28
GPT-5 nano	2025-08	0.13	0.46	0.7 [0.4, 1.0]	0.8 [0.7, 0.9]	0.66 [0.62, 0.70]	-0.35 [-0.45, -0.24]	0.01

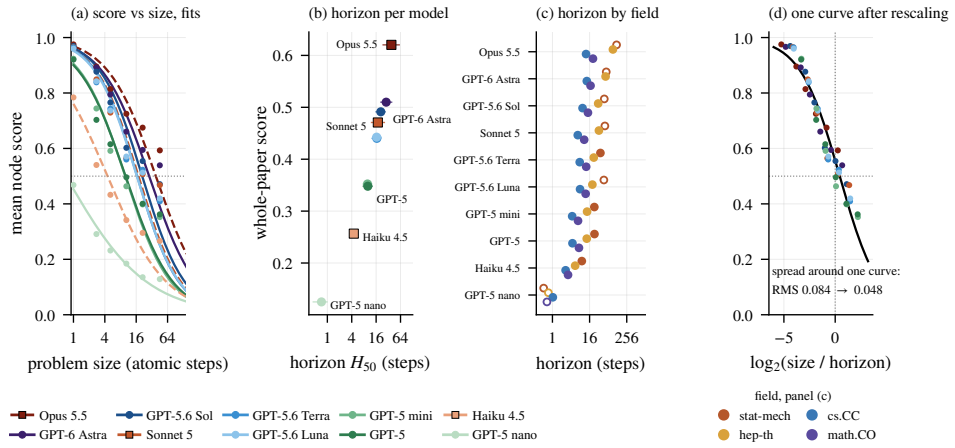

Figure 3: **Measurement 1: score against problem size.** (a) Mean score by problem size (dots) and the fitted logistic curve for each model (lines); a model's research horizon is where its curve crosses one half (dotted line). (b) Research horizon against whole-paper score, with 95% intervals (squares: Claude; circles: OpenAI). (c) Research horizon fitted separately in each field; hollow markers lie beyond the largest problem observed in that field. (d) The binned means of the eight capable models, with problem size divided by each model's horizon, around one common curve (black).

We call the eight models with step scores above 0.9 the *capable* models, and the six strongest of them (Opus 5.5, GPT-6 Astra, the three GPT-5.6 models and Sonnet 5) the *strong* models. Among the capable models, the gap between whole papers and steps, not step accuracy, separates them. Opus 5.5 scores highest on whole papers, ahead of Sonnet 5 on 19 of 20 papers; the next five models are close to each other, and the four smallest or oldest models are clearly behind, also when any one field is removed (Table 1).

4.2 Measurement 1: The Research Horizon

Scores fall smoothly with problem size, and one logistic curve per model describes them well (Figure 3a). The curves have similar slopes for the capable models; what differs mainly is how far to the right each one sits. That position is the research horizon: about 40 steps for Opus 5.5, 28 for GPT-6 Astra, 17 to 21 for the GPT-5.6 models and Sonnet 5, 10 for GPT-5 and GPT-5 mini, 4.5 for Haiku 4.5, and below one step for GPT-5 nano (Table 1). Although fitted on all problems, the horizon ranks models almost exactly as the whole-paper score does (Figure 3b).

Fields differ far more at the scale of papers than at the scale of steps (Figure 3c). Within the observed range of problem sizes, capable models reach problems about two to seven times larger in high-energy theory than in combinatorics and complexity (Opus 5.5: about 90 steps against 21 and 12); in statistical mechanics the strongest models stay above one half on every problem, the largest of which has 38 steps, so their fitted horizons there are extrapolations (Appendix H). Complexity papers still get higher whole-paper scores than combinatorics papers because their trees are about half as large; per step of size they are the hardest. Single steps are solved almost equally well in all fields, so the difference lies in chaining them, which suggests that steps in a proof are harder to build on than steps in a physics derivation.

Measured in units of its own horizon, every capable model follows approximately the same curve (Figure 3d). Rescaling problem size by each model's horizon halves the spread of the data around a single logistic, and the same holds for each model within each field; the remaining small differences in slope are resolved by a finer regression (Appendix K). To a good approximation, one number, the horizon, summarises a capable model in a field.

The horizon grows across the leading releases (Figure 1c). Across the top-tier releases of each lab (GPT-5, GPT-5.6 Sol, GPT-6 Astra and Opus 5.5) it doubles about every eight months (8.2 with Sonnet 5, which also set a new best), numerically close to the seven months reported for software tasks (Kwa et al., 2025); over 13.5 months this is a trend rather than a law, and without GPT-5 it is about three months (Appendix H).

4.3 Measurement 2: recovering a paper part by part

Giving a model more of the paper's structure raises its score steadily (Figure 4a). The curves of the capable models agree once the paper is broken into small pieces and separate as the pieces grow, and they differ between fields (Figure 4b).

The decomposition deficit summarises how much of a paper a model still misses, averaged over every level of decomposition from the whole paper to its steps (Figure 4d). Opus 5.5 still misses about a ninth of the paper, the other strong models an eighth to a sixth, GPT-5 and GPT-5 mini a quarter, Haiku 4.5 two fifths and GPT-5 nano two thirds. The fields fall in the same order as for the horizon: statistical-mechanics papers leave the least missing and complexity papers the most.

What models miss shrinks across the leading releases (Figure 4f). Across the same four releases as for the horizon, the decomposition deficit halves about once a year, the mirror image of the growing horizon, again a trend over four releases rather than a law (Appendix I).

The strong models recover their gap in the same way (Figure 4c). Relative to each model's own gap, a quarter of the tree recovers about 60% and half the tree over 80%, with little variation between the six strongest models; weaker models recover more slowly.

The first split matters most for the strong models (Figure 4e). Splitting a paper once recovers about 40 to 45% of the gap for Opus 5.5, Astra and the GPT-5.6 models (35% for Sonnet 5) but 30% or less for the four weakest models, which gain most at the second level; since a split also supplies premises, this measures the value of the paper's structure rather than planning ability alone.

The gap opens where steps are combined: for the strong models, more than four in five parent problems score below their weakest child, and splitting a problem almost never lowers the score (Appendix I).

4.4 Failure modes on large problems

Both measurements show that models can do the steps of a paper but lose score when the steps have to be put together. A low score on a large problem may mean a full derivation with errors or a correct but partial answer; the difference matters for what the horizon measures and how to raise it, since the first calls for more reliable steps and the second for carrying a long argument through to the end.

The deficit lies beyond the steps. Score falls steadily with problem size for every model, without a size at which it collapses (Figure 5a). Splitting each model's whole-paper deficit relative to Opus 5.5 into weaker steps and weaker composition shows that over 90% of it comes from composition for the strong models and about 80% for GPT-5 and GPT-5 mini; only the two weakest models fall short

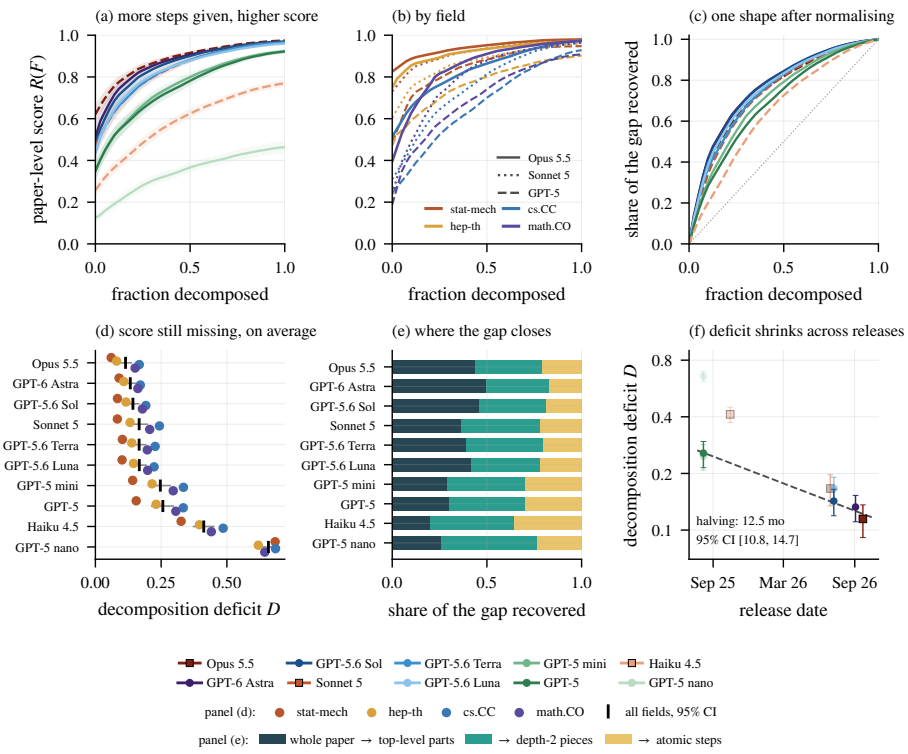

Figure 4: **Measurement 2: score against decomposition.** (a) Decomposition curves: the paper-level score as the paper is posed in finer pieces, from the whole paper ($x = 0$) to all atomic steps ($x = 1$); mean of 20 papers with standard error. (b) The same curves by field for three models. (c) Each curve rescaled to the share of that model's gap between whole paper and steps; the strong models share one shape. (d) Decomposition deficit D, the score still missing averaged over the curve in (a), by field (dots) and over all papers (black ticks, with 95% intervals). (e) Share of the gap recovered by splitting the paper into its top-level parts, then into second-level pieces, then into atomic steps. (f) Decomposition deficit against release date, with 95% intervals; the dashed line is the trend over the four leading releases of Figure 1c.

mainly on the steps themselves, Haiku 4.5 for more than half and GPT-5 nano for all of its deficit (Figure 5b). The split is an accounting identity whose remainder includes what a split supplies, premises and room to answer, as well as combining steps; answer length does not explain it, since models write longer answers for larger problems and longer answers do not score higher at a given size (Appendix K).

Models leave work out rather than get it wrong. We classified the judge's explanations for a stratified sample of 700 judgments (Appendix K). Among low scores, omitted results, omitted derivations and answers to a narrower question account for about three quarters of the lost score, and no error at all is identified in about four fifths; explicit errors account for about a sixth, from 14% at the root to 23% at depth three or more, a difference the sample does not resolve (Figure 5d). These are the judge's diagnoses: it grades against the step's reference, without the larger problem's statement, and was not checked by experts. A small share of low scores at the root are correct proofs by a route other than the paper's. These labels describe what the judge reports, not independently verified correctness.

Progress comes from below. In the release pairs shown, newer models fix almost all of the steps that older models failed, but only a minority of their failures on large problems (Figure 5c). Most measured progress is therefore on problems of moderate size, and completing long derivations is where most of the remaining gap lies.

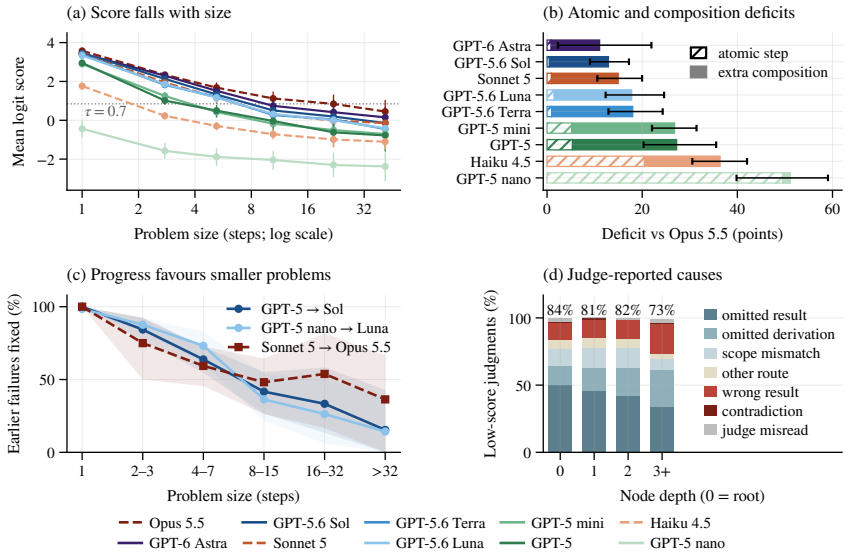

Figure 5: **Failure modes on large problems.** (a) Mean logit of the node score (clipped to $[0.02, 0.98]$) by problem size, with 95% intervals; the dotted line marks a score of 0.7. (b) Each model's whole-paper deficit relative to Opus 5.5, split into the part explained by weaker step scores (hatched) and the remaining gap between steps and whole paper (solid). (c) Share of an earlier model's failed problems that a later model solves, by problem size. (d) Main cause of a low score according to the judge's rationale, by depth in the tree; numbers above the bars give the share of lost score with no error identified. Bars and bands in (b) and (c) are 95% paper-bootstrap intervals; details in Appendix K.

Checks on the measurements. A judge from another lab agrees closely with ours, and the three disputed root problems explain most of Opus 5.5's lead over GPT-6 Astra but not the gap between steps and whole papers (Appendix C).

5 Limitations

A paper records the final derivation, not the exploration and failed attempts behind it (Liu et al., 2026b), so the horizon measures how much of a written argument a model can reconstruct, and the tree is the builder's reconstruction of that argument, a compromise we accept because papers are the only complete, public record of research at scale. An LLM judge grades against the paper's own route, with references not checked by experts and trees from one builder family that we did not rebuild. Twenty theoretical papers do not represent research as a whole, the release trends rest on four models, and splitting a paper also gives sub-problems premises and more room to answer (Appendix B).

6 Conclusion

We asked how much of existing research a language model can reconstruct on its own. On recent theoretical papers, today's models solve almost every individual step but recover much less of a paper as a whole, and the best of them handles problems of about 40 steps. That horizon has grown quickly across recent releases, and when models fall short on a large problem they mostly leave parts of the argument out rather than get them wrong. Because ResearchTree needs nothing but a paper, the measurement can be rebuilt as new papers and models appear; checking the references with experts and testing whole-paper answers with a matched length budget are the most important next steps.

AI USE STATEMENT

In this work, we used generative AI tools for generating data, designing methodology and experiments, implementing methods and interpreting results. Generating the evaluation data is the method itself: an LLM builds the trees and reference solutions, the models under test are LLMs, and LLMs judge the answers (Sections 2 and 3). Beyond this, LLM coding agents working under the authors' direction proposed and implemented analyses (for example the decomposition deficit and its comparison with thresholded summaries, Appendix I), wrote the analysis and figure code, and drafted interpretations of results, which the authors checked against the data. Formulating mathematical claims and assisting with proofs are not applicable to this work: its mathematical content is the evaluated papers' own results, which we do not claim. Additionally, we used generative AI tools for drafting and editing text, writing code, searching the literature, formatting references, and producing the conceptual illustrations in Figures 2 and 8 with an image-generation model. We have reviewed all AI-assisted work: every quantitative claim is traced to a saved analysis script and retained data, all plots are regenerated by scripts from saved results, the illustrations were checked against the protocol and show no data, and every reference was checked against its arXiv or Crossref record. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The study involves no human subjects and no personal data. The trees are derived from publicly posted arXiv papers; the released dataset links to each paper and contains problems and reference solutions derived from it rather than the papers' full text, and we will honour each paper's license. The evaluated models include three from the same family as the builder and the judge; we disclose this, test the judge with a model from another lab (Appendix C), and discuss the remaining risk of self-preference in Section 5. Better measurement of research ability may accelerate automated research, with benefits and risks that extend beyond this evaluation; the dataset itself contains only published theoretical results.

REPRODUCIBILITY STATEMENT

Appendix E specifies the evaluation and decomposition-curve algorithms; Appendix F provides the prompts, and the remaining appendices document the trees, model settings, retry policy and provenance. We retain the problem statements, reference solutions, answers (including superseded attempts), judgments and analysis scripts, so the reported statistics can be recomputed without new model calls. We also release the campaign configuration (model identifiers, harness versions, reasoning settings, answer and time limits) and the outputs of the paper filter (Appendix G). We release these materials as a public dataset, together with the code that builds trees from new papers, so that the evaluation can be renewed and extended to other fields; the link is withheld for anonymous review.

REFERENCES

Sungeun An, Swanand Ravindra Kadhe, Shailja Thakur, Chad DeLuca, and Hima Patel. STaD: Scaffolded Task Design for Identifying Compositional Skill Gaps in LLMs. *arXiv preprint arXiv:2604.18177*, 2026.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring Length Generalization in Large Language Models. In *Advances in Neural Information Processing Systems 35*, 2022. doi: 10.52202/068431-2793.

Ali Ansari, Haoran Sun, Andy Zeyi Liu, Mark Jabbour, Yongshan Ding, Steven Girvin, Yu He, Sohrab Ismail-Beigi, Aleksander Kubica, Owen D. Miller, Corey O'Hern, Vidvuds Ozolins, David Poland, A. Douglas Stone, Frank C. van den Bosch, Logan Wright, Navid Akbari, Santanu Antu, Kangle Cai, Andrew Calabrese-Day, Mateo Crdenes Wuttig, Meng Cheng, Barry T. Chiang, Ali Ghorashi, Shouzhen Gu, Haoyang Huang, Zhibo Kang, Lukas Kienesberger, Hantian Liu, Charles Lomba, Zhongling Lu, Wenchao Ma, Rohin E. McIntosh, Evan McKinney, Ivan

Rojkov, Xulei Sun, Yarone Meir Tokayer, Naveen Balaji Umasankar, Mira Varma, Leda Wang, Qimin Wang, Tyler Wang, Haoyu Wei, Jinming Yang, Jinchen Zhao, Sherlock Tingrui Zhao, Qinyuan Zheng, Jay S. Zou, Lucas Baker, Arman Cohan, and John Sous. How Good Are Frontier Models at Physics? Expert Re-Grading Reveals Broken Evaluations and Near-Saturation of Leading Benchmarks. *arXiv preprint arXiv:2609.13009*, 2026.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quionero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating Large Language Models Towards Improved Human Health. *arXiv preprint arXiv:2505.08775*, 2025.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. ProofNet: Autoformalizing and Formally Proving Undergraduate-Level Mathematics. *arXiv preprint arXiv:2302.12433*, 2023.

Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. In *International Conference on Machine Learning (ICML)*, 2024.

Mislav Balunovi, Jasper Dekoninck, Ivo Petrov, Nikola Jovanovi, and Martin Vechev. MathArena: Evaluating LLMs on Uncontaminated Math Competitions. *arXiv preprint arXiv:2505.23281*, 2025.

Nitay Calderon, Roi Reichart, and Rotem Dror. The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs. *arXiv preprint arXiv:2501.10970*, 2025.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mdry. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. *arXiv preprint arXiv:2410.07095*, 2024.

Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025a.

Kevin Chen, Marco Cusumano-Towner, Brody Huval, Aleksei Petrenko, Jackson Hamburger, Vladlen Koltun, and Philipp Krhenbhl. Reinforcement Learning for Long-Horizon Interactive LLM Agents. *arXiv preprint arXiv:2502.01600*, 2025b.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A Theorem-driven Question Answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7889–7901, 2023. doi: 10.18653/v1/2023.emnlp-main.489.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the Surface: Measuring Self-Preference in LLM Judgments. *arXiv preprint arXiv:2506.02592*, 2025c.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. *arXiv preprint arXiv:2410.05080*, 2024.

Daniel J. H. Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Mnchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. Theoretical Physics Benchmark (TPBench) – a Dataset and Study of AI Reasoning Capabilities in Theoretical Physics. *arXiv preprint arXiv:2502.15815*, 2025.

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

Vincent Cohen-Addad, Dimitris Paparas, Ernest van Wijland, Max Springer, Julien Canitrot-Paradis, Honghao Lin, David Woodruff, Adarsh Kumarappan, Rajesh Jayaram, Rudrajit Das, Lalit Jain, Ola Svensson, Silvio Lattanzi, Mislav Balunovic, Theophane Weber, and Vahab Mirrokni. TCS-BENCH: Benchmarking State-of-the-Art Generative AI Theoretical Computer Science Research Ability. *arXiv preprint arXiv:2608.09538*, 2026.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 645:633–638, 2025.

Jasper Dekoninck, Nikola Jovanovi, Tim Gehrunger, Kri Rgnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond Benchmarks: MathArena as an Evaluation Platform for Mathematics with LLMs. *arXiv preprint arXiv:2605.00674*, 2026.

Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, et al. SWE-Bench Pro: Can AI Agents Solve Long-Horizon Software Engineering Tasks? *arXiv preprint arXiv:2509.16941*, 2025.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. Successive Prompting for Decomposing Complex Questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.18653/v1/2022.emnlp-main.81.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and Fate: Limits of Transformers on Compositionality. *arXiv preprint arXiv:2305.18654*, 2023.

Damon Falck, Samer Sabri, Anja Surina, Thom Foster, Anya Sims, Sam Devlin, Dylan Rogers, Tantum Collins, Kaloyan Aleksiev, Louis Kirsch, and Edward Hughes. Training AI Scientists to Replicate Research. *arXiv preprint arXiv:2608.13331*, 2026.

Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael P. Brenner. HARDMath: A Benchmark Dataset for Challenging Problems in Applied Mathematics. In *International Conference on Learning Representations (ICLR)*, 2025.

Moran Feldman and Amin Karbasi. Gödel Test: Can Large Language Models Solve Easy Conjectures? *arXiv preprint arXiv:2509.18383*, 2025.

Moritz Firsching, Paul Lezeau, Salvatore Mercuri, Miklós Z. Horváth, Yaël Dillies, Calle Sönne, Eric Wieser, Fred Zhang, Thomas Hubert, Blaise Agüera y Arcas, and Pushmeet Kohli. Formal Conjectures: An Open and Evolving Benchmark for Verified Discovery in Mathematics. *arXiv preprint arXiv:2605.13171*, 2026.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-MATH: A Universal Olympiad Level Mathematic Benchmark For Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2025.

Suhaas Garre, Erik Knutsen, Sushant Mehta, and Edwin Chen. Riemann-Bench: A Benchmark for Moonshot Mathematics. *arXiv preprint arXiv:2604.06802*, 2026.

Haosen Ge, Hamsa Bastani, and Osbert Bastani. Are AI Capabilities Increasing Exponentially? A Competing Hypothesis. *arXiv preprint arXiv:2602.04836*, 2026.

Hejia Geng, Zesen Huang, Haoyang Li, Wenbin Li, Koutian Wu, Zihan Zhou, Yuanbo Pang, Weihao Liu, Zigong Xu, Zhiping Li, Zongzheng Zhang, Chuanfei Dong, Jiankai Sun, Tianzhe Zheng, Fengyu Xie, Yue Ma, Yueheng Shi, Tong Xie, Zonglin Di, Xianrong Liu, Qucheng Gao, Yimin Liu, Jiaming Pan, Sheng Huang, Xiao-Han Ma, Lanqing Yuan, Zhenlin Zhu, Ziang Liu, Ziyang Xu, Junkai Wang, Kangkai Liang, Jiayi Xian, Zehong Zhao, Liuwei Xu, Jingxu Xie, Peijin Zhang, Qiang Gao, Chengyi Xing, Zhe Zhao, Xi Wang, Yaopeng Xing, Xing Meng, Zhenfei Yin, Yingcheng Wu, and Ling Yang. ScienceIDE: Turning World's Scientific Codebase into Agent Learnable Environments. *arXiv preprint arXiv:2609.19134*, 2026.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Jrviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024.

Shashwat Goel, Rishi Hazra, Dulhan Jayalath, Timon Willi, Parag Jain, William F. Shen, Ilias Leontiadis, Francesco Barbieri, Yoram Bachrach, Jonas Geiping, and Chenxi Whitehouse. Training AI Co-Scientists Using Rubric Rewards. *arXiv preprint arXiv:2512.23707*, 2025.

Shahriar Golchin and Mihai Surdeanu. Time Travel in LLMs: Tracing Data Contamination in Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024.

Santiago Gonzalez, Alireza Amiri Bavandpour, Peter Ye, Edward Zhang, et al. QEDBench: Quantifying the alignment gap in automated evaluation of university-level mathematical proofs. *arXiv preprint arXiv:2602.20629*, 2026.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, Jos R Penads, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Dewi Gould, Francis Rhys Ward, Anders Cairns Woodruff, Rauno Arike, Josh Hills, Alex Serrano, Ida Caspary, Jason Ross Brown, Jo J. Jiao, Patrick Leask, Twm Stone, Ram Potham, Ionut Gabriel Stan, Harry Mayne, Simeon Hellsten, Shubhorup Biswas, Ariana Azarbal, William L. Anderson, Elle Najt, Ryan Greenblatt, and Julian Stastny. Think Fast: Estimating No-CoT Task-Completion Time Horizons of Frontier AI Models. *arXiv preprint arXiv:2606.07157*, 2026.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher R, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *arXiv preprint arXiv:2308.11462*, 2023.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as Rewards: Reinforcement Learning Beyond Verifiable Domains. In *International Conference on Learning Representations (ICLR)*, 2026.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024. doi: 10.18653/v1/2024.acl-long.211.

Linyang He, Qiyao Yu, Hanze Dong, Baohao Liao, Xinxing Xu, Micah Goldblum, Jiang Bian, and Nima Mesgarani. LiveMathematicianBench: A Live Benchmark for Research-Level Mathematical Reasoning with Proof Sketches. *arXiv preprint arXiv:2604.01754*, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Anson Ho, Jean-Stanislas Denain, David Atanasov, Samuel Albanie, and Rohin Shah. A Rosetta Stone for AI Benchmarks. *arXiv preprint arXiv:2512.00193*, 2025.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A. Smith. Fluid Language Model Benchmarking. *arXiv preprint arXiv:2509.11106*, 2025.

Arian Hosseini, Alessandro Sordoni, Daniel Toyama, Aaron Courville, and Rishabh Agarwal. Not All LLM Reasoners Are Created Equal. *arXiv preprint arXiv:2410.01748*, 2024.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? In *Conference on Language Modeling (COLM)*, 2024.

Yichen Huang and Lin F. Yang. Winning Gold at IMO 2025 with a Model-Agnostic Verification-and-Refinement Pipeline. *arXiv preprint arXiv:2507.15855*, 2025.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, et al. Reinforcement Learning with Rubric Anchors. *arXiv preprint arXiv:2508.12790*, 2025.

Thomas Hubert, Rishi Mehta, Laurent Sartran, Miklós Z. Horváth, Goran Žužić, Eric Wieser, Aja Huang, Julian Schrittwieser, Yannick Schroecker, Hussain Masoom, Ottavia Bertolli, Tom Zahavy, Amol Mandhane, Jessica Yung, Iuliya Beloshapka, Borja Ibarz, Vivek Veeriah, Lei Yu, Oliver Nash, Paul Lezeau, Salvatore Mercuri, Calle Sönne, Bhavik Mehta, Alex Davies, Daniel Zheng, Fabian Pedregosa, Yin Li, Ingrid von Glehn, Mark Rowland, Samuel Albanie, Ameya Velingker, Simon Schmitt, Edward Lockhart, Edward Hughes, Henryk Michalewski, Nicolas Sonnerat, Demis Hassabis, Pushmeet Kohli, and David Silver. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651(8106):607–613, 2025. doi: 10.1038/s41586-025-09833-y.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. In *International Conference on Learning Representations (ICLR)*, 2025a.

Naman Jain, Jaskirat Singh, Manish Shetty, Liang Zheng, Koushik Sen, and Ion Stoica. R2E-Gym: Procedural Environments and Hybrid Verifiers for Scaling Open-Weights SWE Agents. In *Conference on Language Modeling (COLM)*, 2025b.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DiscoveryWorld: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Albert Q. Jiang, Sean Welleck, Jin Peng Zhou, Wenda Li, Jiacheng Liu, Mateja Jamnik, Timothe Lacroix, Yuhuai Wu, and Guillaume Lample. Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs. *arXiv preprint arXiv:2210.12283*, 2022.

Bo Jiang. HindSight: Evaluating LLM-Generated Research Ideas via Future Impact. *arXiv preprint arXiv:2603.15164*, 2026.

Xitai Jiang, Zihan Tang, Wenze Lin, Yang Yue, Shenzhi Wang, and Gao Huang. From Reasoning Chains to Verifiable Subproblems: Curriculum Reinforcement Learning Enables Credit Assignment for LLM Reasoning. *arXiv preprint arXiv:2605.22074*, 2026.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? In *International Conference on Learning Representations (ICLR)*, 2024.

Sayash Kapoor, Peter Kirgis, Andrew Schwartz, Stephan Rabanser, J. J. Allaire, Rishi Bommasani, Harry Coppock, Magda Dubois, Gillian K Hadfield, Andrew B. Hall, et al. Open-World Evaluations for Measuring Frontier AI Capabilities. *arXiv preprint arXiv:2605.20520*, 2026.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed Prompting: A Modular Approach for Solving Complex Tasks. *arXiv preprint arXiv:2210.02406*, 2022.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. doi: 10.18653/v1/2021.naacl-main.324.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. In *International Conference on Learning Representations (ICLR)*, 2024a.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024b.

Miyoung Ko, Sue Hyun Park, Joonsuk Park, and Minjoon Seo. Hierarchical deconstruction of LLM reasoning: A graph-based framework for analyzing knowledge utilization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Chris Painter, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI Ability to Complete Long Software Tasks. *arXiv preprint arXiv:2503.14499*, 2025.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. In *Conference on Language Modeling (COLM)*, 2025.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. LAB-Bench: Measuring Capabilities of Language Models for Biology Research. *arXiv preprint arXiv:2407.10362*, 2024.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025a.

Xiang Lisa Li, Farzaan Kaiyom, Evan Zheran Liu, Yifan Mai, Percy Liang, and Tatsunori Hashimoto. AutoBencher: Towards Declarative Benchmark Construction. In *International Conference on Learning Representations (ICLR)*, 2025b.

Zongxia Li, Zhongzhi Li, Yucheng Shi, Ruhan Wang, Junyao Yang, Zhichao Liu, Xiyang Wu, Anhao Li, Yue Yu, Ninghao Liu, et al. Long-Horizon-Terminal-Bench: Testing the Limits of Agents on Long-Horizon Terminal Tasks with Dense Reward-Based Grading. *arXiv preprint arXiv:2607.08964*, 2026.

Xiao Liang, Zhong-Zhi Li, Zhenghao Lin, Eric Hancheng Jiang, Hengyuan Zhang, Yelong Shen, Kai-Wei Chang, Ying Nian Wu, Yeyun Gong, and Weizhu Chen. Training LLMs for Divide-and-Conquer Reasoning Elevates Test-Time Scalability. *arXiv preprint arXiv:2602.02477*, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. ZebraLogic: On the Scaling Limits of LLMs for Logical Reasoning. In *International Conference on Machine Learning (ICML)*, 2025.

Jiachen Liu, Maestro Harmon, and Zechen Zhang. Sci-reasoning: A dataset decoding ai innovation patterns. *arXiv preprint arXiv:2601.04577*, 2026a.

Jiachen Liu, Jiaxin Pei, Jintao Huang, Chenglei Si, Ao Qu, Xiangru Tang, et al. The last human-written paper: Agent-native research artifacts. *arXiv preprint arXiv:2604.24658*, 2026b.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 2024. doi: 10.1162/tacl_a_00638.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. doi: 10.18653/v1/2023.emnlp-main.153.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. ResearchBench: Benchmarking LLMs in Scientific Discovery via Inspiration-Based Task Decomposition. *arXiv preprint arXiv:2503.21248*, 2025.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Haotian Luo, Huaisong Zhang, Xuelin Zhang, Haoyu Wang, Zeyu Qin, Wenjie Lu, Guozheng Ma, Haiying He, Yingsha Xie, Qiyang Zhou, et al. UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios. *arXiv preprint arXiv:2509.21766*, 2025.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve Mathematical Reasoning in Language Models by Automated Process Supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Jicheng Ma, Guohua Wang, Xinhua Feng, Yiming Liu, Zhichao Hu, and Yuhong Liu. EternalMath: A Living Benchmark of Frontier Mathematics that Evolves with Human Discovery. *arXiv preprint arXiv:2601.01400*, 2026.

Wenjie Ma, Andrei Cojocaru, Neel Kolhe, Bradley Louie, Robin Said Sharif, Haihan Zhang, Vincent Zhuang, Matei Zaharia, and Sewon Min. Reliable Fine-Grained Evaluation of Natural Language Math Proofs. *arXiv preprint arXiv:2510.13888*, 2025.

Hamed Mahdavi, Pouria Mahdavinia, Samira Malek, Pegah Mohammadipour, Alireza Hashemi, Majid Daliri, Alireza Farhadi, Amir Khasahmadi, Niloofar Mireshghallah, and Vasant Honavar. RefGrader: Automated Grading of Mathematical Competition Proofs using Agentic Workflows. *arXiv preprint arXiv:2510.09021*, 2025a.

Sadegh Mahdavi, Branislav Kisacanin, Shubham Toshniwal, Wei Du, Ivan Moshkov, George Armstrong, Renjie Liao, Christos Thrampoulidis, and Igor Gitman. Scaling Generative Verifiers For Natural Language Mathematical Proof Verification And Selection. *arXiv preprint arXiv:2511.13027*, 2025b.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, et al. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces. In *International Conference on Learning Representations (ICLR)*, 2026.

METR. How does time horizon vary across domains? https://metr.org/blog/2025-07-14-how-does-time-horizon-vary-across-domains/, 2025. Blog post, 14 July 2025.

Grgoire Mialon, Clmentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for General AI Assistants. In *International Conference on Learning Representations (ICLR)*, 2024.

Tingjia Miao, Wenkai Jin, Muhua Zhang, Jinxin Tan, Yuelin Hu, Tu Guo, Jiejun Zhang, Yuhan Wang, Wenbo Li, Yinuo Gao, Shuo Chen, Weiqi Jiang, Yayun Hu, Zixing Lei, Xianghe Pang, Zexi Liu, Yuzhi Zhang, Linfeng Zhang, Kun Chen, Wei Wang, Weinan E, and Siheng Chen. PRL-Bench: A Comprehensive Benchmark Evaluating LLMs' Capabilities in Frontier Physics Research. *arXiv preprint arXiv:2604.15411*, 2026.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2025.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. SWE-Lancer: Can Frontier LLMs Earn $1 Million from Real-World Freelance Software Engineering? In *International Conference on Machine Learning (ICML)*, 2025.

Shambhavi Mishra, Gaurav Sahu, Marco Pedersoli, Laurent Charlin, Jose Dolz, and Christopher Pal. AInstein: Can LLMs Solve Research Problems From Parametric Memory Alone? *arXiv preprint arXiv:2510.05432*, 2025.

Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C. Landsness, Daniel L. Barabasi, Siddharth Narayanan, Nicky Evans, Shriya Reddy, Martha Foiani, Aizad Kamal, Leah P. Shriver, Fang Cao, Asmamaw T. Wassie, Jon M. Laurent, Edwin Melville-Green, Mayk Caldas, Albert Bou, Kaleigh F. Roberts, Sladjana Zagorac, Timothy C. Orr, Miranda E. Orr, Kevin J. Zwezdaryk, Ali E. Ghareeb, Laurie McCoy, Bruna Gomes, Euan A. Ashley, Karen E. Duff, Tonio Buonassisi, Tom Rainforth, Randall J. Bateman, Michael Skarlinski, Samuel G. Rodriques, Michaela M. Hinks, and Andrew D. White. Kosmos: An AI Scientist for Autonomous Discovery. *arXiv preprint arXiv:2511.02824*, 2025.

Fan Nie, Ken Ziyu Liu, Zihao Wang, Rui Sun, Wei Liu, Weijia Shi, Huaxiu Yao, Linjun Zhang, Andrew Y. Ng, James Zou, Sanmi Koyejo, Yejin Choi, Percy Liang, and Niklas Muennighoff. UQ: Assessing Language Models on Unsolved Questions. *arXiv preprint arXiv:2508.17580*, 2025.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement, Consistency, and Bias. *arXiv preprint arXiv:2606.19544*, 2026.

Alexander Novikov, Ngn V, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving Test Set Contamination in Black Box Language Models. In *International Conference on Learning Representations (ICLR)*, 2024.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13:6793, 2022. doi: 10.1038/s41467-022-34591-0.

Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. Training Software Engineering Agents and Verifiers with SWE-Gym. In *International Conference on Machine Learning (ICML)*, 2025.

Shrey Pandit, Austin Xu, Xuan-Phi Nguyen, Yifei Ming, Caiming Xiong, and Shafiq Joty. Hard2Verify: A Step-Level Verification Benchmark for Open-Ended Frontier Math. *arXiv preprint arXiv:2510.13744*, 2025.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM Evaluators Recognize and Favor Their Own Generations. *arXiv preprint arXiv:2404.13076*, 2024.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simn Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, et al. GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. In *International Conference on Learning Representations (ICLR)*, 2026.

Ethan Perez, Sam Ringer, Kamil Lukoit, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noem Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunovi, Nikola Jovanovi, and Martin Vechev. Proof or Bluff? Evaluating LLMs on 2025 USA Math Olympiad. *arXiv preprint arXiv:2503.21934*, 2025.

Antoine Peyronnet, Fabian Gloeckle, and Amaury Hayat. LemmaBench: A Live, Research-Level Benchmark to Evaluate LLM Capabilities in Mathematics. *arXiv preprint arXiv:2602.24173*, 2026.

QuocViet Pham, Elvir Karimov, Andrey Galichin, Ivan Oseledets, et al. TheoremBench: Evaluating LLMs on theorem proving in formal mathematics. *arXiv preprint arXiv:2606.09450*, 2026.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Sren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gzdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever,

Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Mart Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Vclav Rozho, Vincent Ginis, Christian Stump, Niv Cohen, Rafa Powiata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givr, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar ngquist, Yanxu Chen, Harrison K Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, Jrmy Androletti, Raghav Singhal, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khnh, Paul Rosu, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio, Fereshteh Kazemi, Ziye Chen, Bir Blint, Eve J. Y. Lo, Jiaqi Wang, Maria Inłs S. Nunes, Jeremiah Milbauer, M Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan, Emma Rodman, Jacob Drori, Carl J Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep, Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobc, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Mohammadreza Mofayezi, Alexander Piperski, David K. Zhang, Kostiantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua Duersch, Vage Taamazyan, Andrew Ho, Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta, Edson Oliveira, Joseph W. Jackson, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Bita Golshani, David Stap, Egor Kretov, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegozo Rodriguez, Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik Ekstrm, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peaflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristyy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah Carroll, Andrew R. Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham,

Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiit Yaln, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Bosc, Oleg Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Stefano Cavalleri, Olle Hggstrm, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hernndez-Cmara, Emanuele Rodol, Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Isaac C. McAlister, Alejandro Jos Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel Perry, Luke Askew, Ral Adrin Huerta Rodrguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin, Benjmin Borbs, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran uc Huy, Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber, Warren S. Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub ucki, Russell Campbell, Asankhaya Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gunjan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky, Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright, Sergei Bogdanov, Niels Mndler, Sren Mller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mtys Vincze, Siranut Usawasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin, Jiang Muzhen, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin, Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P, Michael Richmond, Joseph McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubi, Samuele Sala, Stephen Ebert, Jean Kaddour, Manuel Schottdorf, Dianzhuo Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved, Ali ElSheikh, S Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilm Zouhar, Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N. Nhu, Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok, Aarush Sinha, Gaoxiang Luo, Long Le, Mickal Noy, Micha Perekiewicz, Ioannis Pantidis, Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata, Edoardo M. Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang, Anupam Nayak, Diogo M. Caetano, Antonio A. W. L. Wong, Maria del Rio-Chanona, Dniel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Gins, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna,

Sina Rismanchian, Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J. Overholt, Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony CY Pang, Adam Wecker, Yifan Xiong, Fanfei Li, Lukas S. Huber, Joshua Jaeger, Romano De Maddalena, Xing Han L, Yuhui Zhang, Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu Zhang, Ankit Agrawal, Li S. Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briaski, Alessandro Stolfo, Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Shabani, Felix Juefei-Xu, Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven, Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos, Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee, Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamourn, Shuyu Wu, Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi, Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera Marjanovi, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E. Zhang, Zhun Wang, Gal Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee, Zhe Ye, Stefano Ermon, Ignacio D. Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park, Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhanany, Han Lin, Philipp D. Siedler, Jichao Fang, Ritwik Mishra, Kroly Zsolnai-Fehr, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin, Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster, Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben Racz, Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V. Gonalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Ycel, Brandon Christof, Veerupaksh Singla, Zijian Song, Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q. Ma, Joshua Mak, Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha, Qiutong Men, Jiaxuan Wu, Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung, Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E. Chen, Woongyeong Yeo, Eeshaan Jain, Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan, Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu, Samaksh Gulati, Songyang Zhang, Peter Turchin, Christopher W. Bartlett, Christopher R. Scotese, Phuong M. Cao, Ben Wu, Jacek Karwowski, Davide Scaramuzza, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Ashley Zhang, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Steven Dillmann, Zhengxiang Wang, Junyu Luo, Hugo Lunn, Artem Gazizov, Haitz Sez de Ocriz Borde, Ivan Trus, Morgan Hervault, Zheyu Zhang, Bo Chen, Yuchen Wu, Christopher J. Cordier, Gn Kaynar, Cansin Ayvaz, Polina Avdiunina, Johannes Brust, Xingjian Diao, K. D. Meaney, Yifan Gu, Chenyu Wang, Chenzhuo Dong, William Wright, Simon Brave, Owen Root, Jiayuan Liu, Chow Chun Lok, Tianqin Li, Shiyi Du, Dailan He, Lufeiya Liu, Sina Jamalzadegan, Anil Ramakrishna, Xuanqing Xu, Xin Qing, Xin Luo, Wenkai Li, Shi Bo, Filipp Gusev, Maximos Skandalis, Desheng Ma, Chunhui Zhang, Haoran Qiu, Allen G Hart, Rickard Brel Gabrielsson, Ido Akov, Artem Lukoianov, Summer Yue, Alexandr Wang, and Dan Hendrycks. Humanity's Last Exam. *arXiv preprint arXiv:2501.14249*, 2025.

Auguste Poiroux, Antoine Bosselut, and Viktor Kunčak. RLMEval: Evaluating research-level neural theorem proving. *arXiv preprint arXiv:2510.25427*, 2025.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and Narrowing the Compositionality Gap in Language Models. *arXiv preprint arXiv:2210.03350*, 2022.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, Yuku Zhang, Boxuan Jing, Xianqi Yin, Yutong Ren, Zizhuo Fu, Jiaming Ji, Weike Wang, Xudong Tian, Anqi Lv, Laifu Man, Jianxiang Li, Feiyu Tao, Qihua Sun, Zhou Liang, Yushu Mu, Zhongxuan Li, Jing-Jun Zhang, Shutao Zhang, Xiaotian Li, Xingqi Xia, Jiawei Lin, Zheyu Shen, Jiahang Chen, Qiuhao Xiong, Binran Wang, Fengyuan Wang, Ziyang Ni, Bohan Zhang, Fan Cui, Changkun Shao, Qing-Hong Cao, Ming-xing Luo, Yaodong Yang, Muhan Zhang, and Hua Xing Zhu. PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

David Rein, Joel Becker, Amy Deng, Seraphina Nix, Chris Canal, Daniel O'Connel, Pip Arnott, Ryan Bloom, Thomas Broadley, Katharyn Garcia, Brian Goodrich, Max Hasin, Sami Jawhar, Megan Kinniment, Thomas Kwa, Aron Lajko, Nate Rush, Lucas Jun Koba Sato, Sydney Von Arx, Ben West, Lawrence Chan, and Elizabeth Barnes. HCAST: Human-Calibrated Autonomy Software Tasks. *arXiv preprint arXiv:2503.17354*, 2025.

Z. Z. Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, Z. F. Wu, Zhibin Gou, Shirong Ma, Hongxuan Tang, Yuxuan Liu, Wenjun Gao, Daya Guo, and Chong Ruan. DeepSeek-Prover-V2: Advancing Formal Mathematical Reasoning via Reinforcement Learning for Subgoal Decomposition. *arXiv preprint arXiv:2504.21801*, 2025.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation Examples are not Equally Informative: How should that change NLP Leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, 2021. doi: 10.18653/v1/2021.acl-long.346.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational Scaling Laws and the Predictability of Language Model Performance. *arXiv preprint arXiv:2405.10938*, 2024.

Abulhair Saparov and He He. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. *arXiv preprint arXiv:2210.01240*, 2022.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent Laboratory: Using LLM Agents as Research Assistants. *arXiv preprint arXiv:2501.04227*, 2025.

Johannes Schmitt, Gergely Brczi, Jasper Dekoninck, Jeremy Feusi, Tim Gehrunger, Raphael Appenzeller, Pieter Belmans, Alessio Bottini, Jim Bryan, Joo Camarneiro, Ana Cannas da Silva, Niklas Canova, Ana-Maria Castravet, Timo de Wolff, Claudio Fontanari, Filippo Gaia, Baran Hashemi, Daniel Holmes, David Holmes, Aitor Iribar Lopez, Victor Jaeck, Martina Jrgensen, Steven Kelk, Martijn Kool, Stefan Kuhlmann, Adam Kurpisz, Johannes Lengler, Chiara Meroni, Ingmar Metzler, Martin Mller, Samuel Muoz-Echniz, David Muoz-Lahoz, Robert Nowak, Georg Oberdieck, Daniel Platt, Dylan Possama, Gabriel Ribeiro, Aluna Rizzoli, Daria Sakhanda, Ral Snchez Galn, Zheming Sun, Diaaeldin Taha, Josef Teichmann, Richard P. Thomas, Henk van der Pol, Michel van Garrel, Charles Vial, Ignacio Barros, Benjamin Doerr, Peter Grnwald, Henry Liu, David Martins, Aleksandar Mijatovi, Sergej Monavari, Marc Roth, Patrick Schnider, Yannik Schuler, Pim Spelier, Yuuji Tanaka, and Ronald van Luijk. IMProofBench: Benchmarking AI on Research-Level Mathematical Proof Generation. *arXiv preprint arXiv:2509.26076*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-3603.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas. *arXiv preprint arXiv:2506.20803*, 2025.

Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark. *Transactions on Machine Learning Research (TMLR)*, 2024.

Akshit Sinha, Arvindh Arun, Shashwat Goel, Steffen Staab, and Jonas Geiping. The Illusion of Diminishing Returns: Measuring Long Horizon Execution in LLMs. *arXiv preprint arXiv:2509.09677*, 2025.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. PRMBench: A Fine-grained and Challenging Benchmark for Process-Level Reward Models. *arXiv preprint arXiv:2501.03124*, 2025.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's Ability to Replicate AI Research. *arXiv preprint arXiv:2504.01848*, 2025.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Kpf. Reasoning Gym: Reasoning Environments for Reinforcement Learning with Verifiable Rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A Benchmark for Evaluating LLM-based Judges. In *International Conference on Learning Representations (ICLR)*, 2025.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, Shengzhu Yin, Minhui Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, Eliu Huerta, and Hao Peng. SciCode: A Research Coding Benchmark Curated by Scientists. *arXiv preprint arXiv:2407.13168*, 2024.

Qiuyu Tian, Haojie Yin, Yingce Xia, Youyong Kong, and Zequn Liu. ForeSci: Evaluating LLM Agents for Forward-Looking AI Research Judgment. *arXiv preprint arXiv:2606.00644*, 2026.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024. doi: 10.1038/s41586-023-06747-5.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 2022. doi: 10.1162/tacl_a_00475.

George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. PutnamBench: Evaluating Neural Theorem-Provers on the Putnam Mathematical Competition. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. *arXiv preprint arXiv:2404.18796*, 2024.

Haiming Wang, Huajian Xin, Zhengying Liu, Wenda Li, Yinya Huang, Jianqiao Lu, Zhicheng Yang, Jing Tang, Jian Yin, Zhenguo Li, and Xiaodan Liang. Proving Theorems Recursively. In *Advances in Neural Information Processing Systems 37*, 2024a. doi: 10.52202/079017-2753.

Haiming Wang, Huajian Xin, Chuanyang Zheng, Lin Li, Zhengying Liu, Qingxing Cao, Yinya Huang, Jing Xiong, Han Shi, Enze Xie, Jian Yin, Zhenguo Li, Heng Liao, and Xiaodan Liang. LEGO-Prover: Neural Theorem Proving with Growing Libraries. In *International Conference on Learning Representations (ICLR)*, 2024b.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023a. doi: 10.18653/v1/2023.acl-long.147.

Miles Wang, Robi Lin, Kat Hu, Joy Jiao, Neil Chowdhury, Ethan Chang, and Tejal Patwardhan. FrontierScience: Evaluating AI's Ability to Perform Expert-Level Scientific Tasks. *arXiv preprint arXiv:2601.21165*, 2026a.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. *arXiv preprint arXiv:2312.08935*, 2023b.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024c.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D. Goodman. Hypothesis Search: Inductive Reasoning with Language Models. In *International Conference on Learning Representations (ICLR)*, 2024d.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *International Conference on Machine Learning (ICML)*, 2024e.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023c. doi: 10.18653/v1/2023.acl-long.754.

Yuhan Wang, Zhengxi Lu, Yuchen Yan, Kaitao Song, Wenqi Zhang, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. PaperGym: Rubric-Centered Evolution for Research-Plan Generation. *arXiv preprint arXiv:2608.31119*, 2026b.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-Preference Bias in LLM-as-a-Judge. *arXiv preprint arXiv:2410.21819*, 2024.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A Challenging, Contamination-Limited LLM Benchmark. In *International Conference on Learning Representations (ICLR)*, 2025.

Parker Whitfill, Ben Snodin, and Joel Becker. Forecasting AI Time Horizon Under Compute Slowdowns. *arXiv preprint arXiv:2511.19492*, 2025.

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Holden Karnofsky, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes. RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. *arXiv preprint arXiv:2411.15114*, 2024.

Yuhao Wu, Ming Shan Hee, Zhiqing Hu, and Roy Ka-Wei Lee. LongGenBench: Benchmarking Long-Form Generation in Long Context LLMs. In *International Conference on Learning Representations (ICLR)*, 2025.

Yuhuai Wu, Albert Q. Jiang, Wenda Li, Markus N. Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. Autoformalization with Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Hanwen Xing, Pengyun Wang, BingXu Meng, Kumail Alhamoud, Xiang Li, Jicheng Wang, Xin Yu, Xinyang Han, Xiaomin Li, Philip Torr, and Yuexing Hao. CurveShift: Is Agent Progress Scalar? Separating Level from Shape. *arXiv preprint arXiv:2608.00355*, 2026.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark Data Contamination of Large Language Models: A Survey. *arXiv preprint arXiv:2406.04244*, 2024.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, et al. TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025a.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. UGPhysics: A Comprehensive Benchmark for Undergraduate Physics Reasoning with Large Language Models. In *International Conference on Machine Learning (ICML)*, 2025b.

Zhangchen Xu, Junda Chen, Yue Huang, Dongfu Jiang, Jiefeng Chen, Hang Hua, Zijian Wu, Zheyuan Liu, Zexue He, Lichi Li, et al. AutoLab: Can Frontier Models Solve Long-Horizon Auto Research and Engineering Tasks? *arXiv preprint arXiv:2606.05080*, 2026.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. *arXiv preprint arXiv:2504.08066*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint arXiv:2305.10601*, 2023.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. In *International Conference on Learning Representations (ICLR)*, 2025.

Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. LongProc: Benchmarking Long-Context Language Models on Long Procedural Generation. In *Conference on Language Modeling (COLM)*, 2025.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. Skill-Mix: a Flexible and Expandable Family of Evaluations for AI models. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhe Yu, Wenpeng Xing, Yunzhao Wei, Jie Chen, Hongzhi Wang, Xuyang Teng, and Meng Han. Composition Collapse: Stable Factual Knowledge Does Not Imply Compositional Reasoning. *arXiv preprint arXiv:2605.26789*, 2026.

Eric Zelikman, Qian Huang, Gabriel Poesia, Noah D. Goodman, and Nick Haber. Parsel: Algorithmic Reasoning with Language Models by Composing Decompositions. In *Advances in Neural Information Processing Systems 36*, 2023. doi: 10.52202/075280-1367.

Weiqi Zhai, Zhihai Wang, Jinghang Wang, Boyu Yang, Xiaogang Li, Xander Xu, Bohan Wang, Peng Wang, Xingzhe Wu, Anfeng Li, Qiyuan Feng, Yuhao Zhou, Taolin Han, Wenjie Luo, Yiyuan Li, Xiang Zheng, Yaxuan Wang, Ruixiang Luo, Guojie Lin, Peiyao Xiao, Chengliang Xu, Ben Wang, Zeyu Wang, Zichao Chen, Jianan Ye, Yijie Hu, Jialong Chen, Zongwen Shen, Yuliang Xu, An Yang, Bowen Yu, Dayiheng Liu, Junyang Lin, Hu Wei, Que Shen, and Bing Zhao. HLE-Verified: A Systematic Verification and Structured Revision of Humanity's Last Exam. *arXiv preprint arXiv:2602.13964*, 2026.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A Careful Examination of Large Language Model Performance on Grade School Arithmetic. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Jie Zhang, Cezara Petrui, Kristina Nikoli, and Florian Tramr. RealMath: A Continuous Benchmark for Evaluating Language Models on Research-Level Mathematics. *arXiv preprint arXiv:2505.12575*, 2025a.

Linghao Zhang, Shilin He, Chaoyun Zhang, Yu Kang, Bowen Li, Chengxing Xie, Junhao Wang, Maoquan Wang, Yufan Huang, Shengyu Fu, Elsie Nallipogu, Qingwei Lin, Yingnong Dang, Saravan Rajmohan, and Dongmei Zhang. SWE-bench Goes Live! In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025b.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The Lessons of Developing Process Reward Models in Mathematical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025c. doi: 10.18653/v1/2025.findings-acl.547.

Wanjia Zhao, Qinwei Ma, Jingzhe Shi, Shirley Wu, Jiaqi Han, Yijia Xiao, Si-Yuan Chen, Xiao Luo, Ludwig Schmidt, and James Zou. PRISM-Physics: Causal DAG-Based Process Evaluation for Physics Reasoning. *arXiv preprint arXiv:2510.03185*, 2025.

Xueliang Zhao, Wenda Li, and Lingpeng Kong. Decomposing the Enigma: Subgoal-based Demonstration Learning for Formal Theorem Proving. *arXiv preprint arXiv:2305.16366*, 2023.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying Process Errors in Mathematical Reasoning. *arXiv preprint arXiv:2412.06559*, 2024.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. MiniF2F: a cross-system benchmark for formal Olympiad-level mathematics. In *International Conference on Learning Representations (ICLR)*, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.

Denny Zhou, Nathanael Schrli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv preprint arXiv:2205.10625*, 2022.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in Benchmarks? Rethinking Large Language Model Benchmarking with Item Response Theory. *arXiv preprint arXiv:2505.15055*, 2025.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Minghang Zhu, Chuyang Wei, Junhao Xu, Yilin Cheng, Zhumin Chen, and Jiyan He. DEEP-RUBRIC: Evidence-Tree Rubric Supervision for Efficient Reinforcement Learning of Deep Research Agents. *arXiv preprint arXiv:2606.17029*, 2026.

Minhui Zhu, Minyang Tian, Xiaocheng Yang, Tianci Zhou, Lifan Yuan, Penghao Zhu, Eli Chertkov, Shengyan Liu, Yufeng Du, Ziming Ji, Indranil Das, Qingzhi Chen, Junyi Cao, Yufeng Du, Jiabin Yu, Peixue Wu, Jinchen He, Yifan Su, Yikun Jiang, Yujie Zhang, Chang Liu, Ze-Min Huang, Weizhen Jia, Yunkai Wang, Farshid Jafarpour, Yong Zhao, Xinan Chen, Jessie Shelton, Aaron W. Young, John Bartolotta, Wenchao Xu, Yue Sun, Anjun Chu, Victor Colussi, Chris Akers, Nathan Brooks, Wenbo Fu, Jinchao Zhao, Marvin Qi, Anqi Mu, Yubo Yang, Allen Zang, Yang Lyu, Peizhi Mai, Christopher Wilson, Xuefei Guo, Juntai Zhou, Daniel Inafuku, Chi Xue, Luyu Gao, Ze Yang, Yar Hein, Yonatan Kahn, Kevin Zhou, Di Luo, John Drew Wilson, Jarrod T. Reilly, Dmytro Bandak, Ofir Press, Liang Yang, Xueying Wang, Hao Tong, Nicolas Chia, Eliu Huerta, and Hao Peng. Probing the Critical Point (CritPt) of AI Reasoning: a Frontier Physics Research Benchmark. *arXiv preprint arXiv:2509.26574*, 2025.

A Related work

Task horizons and long-horizon agents. Kwa et al. (2025) measure the length of software task, in human working time, that an agent completes half the time, calibrated with timed human baselines (Rein et al., 2025), and find that it doubles about every seven months. Later work applies the idea to other task sets (Gould et al., 2026), uses it for forecasting (Whitfill et al., 2025), questions whether the growth is exponential (Ge et al., 2026), compares horizons across domains and warns against reading them beyond the range of task lengths measured (METR, 2025), separates the level of an agent's success curve from its shape (Xing et al., 2026), and studies how small per-step gains compound over long tasks (Sinha et al., 2025). Long-horizon agent benchmarks cover software engineering (Jimenez et al., 2024; Deng et al., 2025; Miserendino et al., 2025; Merrill et al., 2026), web and computer use (Zhou et al., 2024; Xie et al., 2024; Mialon et al., 2024), tool use with users (Yao et al., 2025), workplace and professional tasks (Xu et al., 2025a; Patwardhan et al., 2026), and very long exploration (Luo et al., 2025; Li et al., 2026); Kapoor et al. (2026) argue for messier, open-world evaluations alongside them. These measure horizon in time or in actions. We measure it in the atomic steps of a written derivation, which needs no human timing, grade coverage rather than pass or fail, and read the whole curve of score against size from a single task.

Research and expert benchmarks. Benchmarks for research engineering ask agents to replicate papers (Starace et al., 2025; Siegel et al., 2024), run ML experiments (Wijk et al., 2024; Chan et al., 2024; Chen et al., 2025a; Xu et al., 2026) or carry out data-driven science (Chen et al., 2024; Laurent et al., 2024; Jansen et al., 2024). Question and problem benchmarks range from school and competition mathematics (Cobbe et al., 2021; Hendrycks et al., 2021; He et al., 2024; Gao et al., 2025) and university science (Wang et al., 2024e; Chen et al., 2023; Xu et al., 2025b; Qiu et al., 2025; Rein et al., 2023; Fan et al., 2025) to research-level physics and computer science (Tian et al., 2024; Chung et al., 2025; Zhu et al., 2025; Miao et al., 2026; Cohen-Addad et al., 2026; Wang et al., 2026a) and research mathematics and expert questions (Glazer et al., 2024; Phan et al., 2025; Petrov et al., 2025; Schmitt et al., 2025; Garre et al., 2026; Feldman & Karbasi, 2025; Nie et al., 2025). Formal benchmarks and provers work in Lean and related systems (Zheng et al., 2022; Azerbayev et al., 2023; Tsoukalas et al., 2024; Firsching et al., 2026; Wu et al., 2022; Trinh et al., 2024; Hubert et al., 2025), and informal systems now reach olympiad gold (Huang & Yang, 2025). Closest to us, several benchmarks draw problems from recent papers: RealMath, LemmaBench, EternalMath and Live-MathematicianBench extract research-level statements from new arXiv papers (Zhang et al., 2025a; Peyronnet et al., 2026; Ma et al., 2026; He et al., 2026), and AInstein asks whether models can solve research problems distilled from papers (Mishra et al., 2025). Several benchmarks already score parts and wholes: PaperBench grades a replication against a rubric tree, SciCode evaluates each subproblem as well as the main problem, RLMEval poses research-level theorems with and without

their auxiliary lemmas (Poiroux et al., 2025), TheoremBench expands theorems into premised variants (Pham et al., 2026), and DepthQA decomposes questions into levels of sub-questions to study how models use knowledge (Ko et al., 2024). These work at one or two fixed levels, chosen by hand or by a dataset's structure. RESEARCHTREE builds the full decomposition of a paper automatically and poses every level as a separate problem, which is what gives a size scale and the two curves, and it covers physics and complexity theory as well as mathematics.

AI research systems. End-to-end research agents generate ideas, run experiments and write papers (Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025; Gottweis et al., 2025; Mitchener et al., 2025; Novikov et al., 2025), and their outputs are judged by reviewers or by the ideas' later impact (Si et al., 2024; 2025; Jiang, 2026; Tian et al., 2026); ResearchBench decomposes discovery into inspiration retrieval and hypothesis steps (Liu et al., 2025). Other work restructures papers for agents: agent-native research artifacts turn a paper into an executable package that keeps the exploration and failures the narrative discards (Liu et al., 2026b), and Sci-Reasoning traces how influential papers build on their predecessors, giving structured reasoning trajectories for training research agents (Liu et al., 2026a). These evaluate or support the research process, but with open-ended outputs whose quality is hard to grade. We evaluate a narrower ability, reconstructing a written derivation, but grade it against a concrete reference at every scale.

Renewable and automatically built evaluations. Benchmarks saturate quickly (Ott et al., 2022; Kiela et al., 2021) and leak into training data (Xu et al., 2024; Golchin & Surdeanu, 2024; Oren et al., 2024; Zhang et al., 2024). Live benchmarks answer this by drawing new items after each model's cutoff (White et al., 2025; Jain et al., 2025a; Zhang et al., 2025b; Balunovi et al., 2025; Dekoninck et al., 2026), and models can write evaluations and hard prompts themselves (Perez et al., 2023; Wang et al., 2023c; Li et al., 2025b;a). Item response theory and related methods estimate item difficulty and model ability on a common scale (Rodriguez et al., 2021; Maia Polo et al., 2024; Hofmann et al., 2025; Zhou et al., 2025; Ho et al., 2025), and observational scaling laws predict capabilities from such scales (Ruan et al., 2024). RESEARCHTREE combines these ideas: a model builds the items, new papers renew them, and problem size in atomic steps gives a difficulty scale that is fixed before any model is tested. The recipe needs only a document that writes out a long derivation, so it could extend to other expert domains built from documents, such as law (Guha et al., 2023).

Decomposition and composition. Models fail to compose skills they have individually: the compositionality gap (Press et al., 2022), chained word problems (Hosseini et al., 2024), skill combinations (Yu et al., 2024), multi-hop questions (Trivedi et al., 2022), perturbed or lengthened problems (Mirzadeh et al., 2025; Anil et al., 2022) and puzzles of growing size (Dziri et al., 2023; Shojaee et al., 2025; Lin et al., 2025) all show accuracy falling as more parts must be combined, and long contexts and long outputs degrade in the same way (Liu et al., 2024; Hsieh et al., 2024; Ye et al., 2025; Wu et al., 2025). Models also make valid single steps but struggle to plan multi-step proofs (Saparov & He, 2022), which some attribute to next-token training (Bachmann & Nagarajan, 2024). Recent studies compare performance on parts and on wholes directly (Yu et al., 2026; An et al., 2026; Zhao et al., 2025). Decomposition is also used to help: prompting methods split problems into sub-questions or plans (Zhou et al., 2022; Khot et al., 2022; Dua et al., 2022; Wang et al., 2023a; Yao et al., 2023; Zelikman et al., 2023; Wang et al., 2024d), and theorem provers decompose goals into lemmas (Jiang et al., 2022; Zhao et al., 2023; Wang et al., 2024b;a; Ren et al., 2025). In RESEARCHTREE the model never sees the tree; decomposition serves to measure, and the decomposition curve quantifies how much a given amount of structure would help. Our finding that parents score below their weakest child extends the compositionality gap from synthetic tasks to research derivations.

Step-level supervision. Process supervision grades the steps within one solution (Uesato et al., 2022; Lightman et al., 2023; Wang et al., 2023b; Luo et al., 2024; Zhang et al., 2025c), with benchmarks for process reward models and step verifiers (Zheng et al., 2024; Song et al., 2025; Pandit et al., 2025). These check a model's own chain of steps. RESEARCHTREE instead asks the model to produce the steps and checks each of its answers against the steps of a fixed reference derivation.

Judges, rubrics and references. LLM judges score free-form answers against criteria or references (Zheng et al., 2023; Liu et al., 2023; Kim et al., 2024a;b), and rubric-based grading scales expert criteria to thousands of items (Arora et al., 2025; Starace et al., 2025). Judges show position, verbosity and self-preference biases (Wang et al., 2024c; Panickssery et al., 2024; Wataoka et al., 2024; Chen et al., 2025c), can be unreliable on hard reasoning (Tan et al., 2025; Norman et al., 2026), are often lenient when grading proofs (Gonzalez et al., 2026), and improve with reference solutions, trained verifiers and panels from several families (Mahdavi et al., 2025a;b; Ma et al., 2025; Verga et al., 2024); references and questions can themselves be wrong (Zhai et al., 2026; Ansari et al., 2026). We grade against references taken from the paper, test the judge with one from another lab, and audit the problems that models disputed.

Training environments for long-horizon reasoning. Reinforcement learning with verifiable rewards elicits long reasoning (Shao et al., 2024; DeepSeek-AI et al., 2025; Lambert et al., 2025), rubrics extend it to answers that cannot be checked automatically (Gunjal et al., 2026; Huang et al., 2025; Goel et al., 2025; Zhu et al., 2026; Wang et al., 2026b), and environment suites supply tasks at scale (Stojanovski et al., 2025; Pan et al., 2025; Jain et al., 2025b; Chen et al., 2025b; Geng et al., 2026; Falck et al., 2026). Curricula of verifiable subproblems and divide-and-conquer training improve long reasoning (Jiang et al., 2026; Liang et al., 2026). The trees we release fit this setting: every node is a graded task with a reference, problem size orders the tasks into a curriculum, and the decomposition curve indicates how much scaffolding to give.

B Extended limitations

A paper is not a research project. A paper records the final derivation, not the work that produced it: the choice of problem, the exploration, the failed attempts and the checks that never appear in print (Liu et al., 2026b). The research horizon therefore measures how much of a known, written-down argument a model can reconstruct, which is a necessary part of research but not all of it. The tree is also a reconstruction: its leaves are the steps a builder reads off the finished paper, not the steps a researcher took, and a model need not cover every leaf to have made real progress. We accept this compromise because the paper is the only record of a research result that is complete, public and available at scale.

Grading against the paper's route. Answers are scored against the paper's own derivation, so a correct proof by another route can score low, and models differ in how often they take other routes or dispute a premise (Appendix C). The cross-lab judge check covers four papers and uses the same reference solutions, so it tests the judge, not the references. Neither judge was calibrated against expert graders (Calderon et al., 2025), LLM judges are often lenient on proofs (Gonzalez et al., 2026), and the classification of judge explanations is itself an LLM assessment. The judge sees the step's problem and reference and the answers along its lineage, but not the statement of the larger problem, so it cannot always tell a justified objection to a misstated larger problem from an unsupported claim; GPT-6 Astra's objection on 2609.04992 is scored as a contradiction of a step that had been given an extra closure premise. All answers along a lineage are scored in one judge call, so the judge could compare the levels with each other; we did not test separate calls per node.

The builder. Builder models from one family (Claude Opus 5.5, and Opus 5 for two trees), the same family as the judge and the top-scoring model, chose every split, weight and leaf, and we did not rebuild trees to measure stability or self-preference. The horizon is therefore measured in this builder's steps, not in human time. Reference solutions include inferred steps and were not checked by domain experts; our audit found two misstated root problems, and errors in a reference affect every score that uses it. Some children state their problem more precisely than their parent, so a high step score does not certify the statement of the root. Premises can also carry a proof route: in the Komlós tree, step `1_2` is given two intermediate results and told which constraints to count and which contradiction to derive. The gap between steps and whole papers therefore measures the value of the whole package a split provides, not the ability to combine steps alone. Finally, a horizon is measured in one builder's steps on one set of trees: trees built from new papers renew the problems, but horizons from different sets of trees are comparable only through models evaluated on both.

Scope and budget. Twenty papers from four theoretical fields do not represent research as a whole: we chose theory because it is cheap to evaluate, and empirical papers, whose steps require running experiments, remain untested. Recent posting does not rule out contamination. The release trends rest on four leading models and change when one is removed, and field horizons beyond the largest problem of a field are extrapolations of the fitted curve (Appendix H). Decomposition changes more than granularity: sub-problems receive premises and together get more room to answer, and we did not test a whole-paper answer with a matching length budget; within the answers we have, length does not explain the size law (Appendix K). Each lab's models ran in their own harness, and the model was instructed, not forced, to stay within its problem. Our intervals reflect the choice of papers, not these systematic factors.

C How far to trust the measurements

Two checks test the grading itself: a judge from another lab re-scores four papers, and an audit examines the root problems that a model disputed. A third paragraph notes how the agent harnesses and failed answers enter the scores.

C.1 A judge from another lab

The main judge, Claude Sonnet 5, shares a family with three of the evaluated models, so we repeated the judging with a model from another lab. On four papers (one per field: 2609.15071, 2609.10823, 2609.04992, 2609.19623) every answer of every model was judged a second time by GPT-5.6 Sol through the same rubric and inputs, giving 3,610 paired (lineage, node) scores. Table 2 gives the per-model comparison; Figure 6 the paired scores and the judge difference by model. The two judges agree across pairs with Pearson 0.945 and Spearman 0.93, and the Sonnet judge scores 0.037 lower than the Sol judge on average; its largest offsets are on GPT-5.6 Luna (-0.054), Terra (-0.053) and Opus 5.5 (-0.052). The two judges rank the models similarly: the rankings correlate at 0.82 on the whole-paper score and at 0.98 on all nodes (Spearman). To test for self-preference we compare, within each paper, how much stricter the Sonnet judge is on Claude models than on OpenAI models. At the weighted root the four per-paper contrasts are -0.004, $+0.009$, -0.031 and -0.032 (mean -0.015, i.e. the Sonnet judge is, if anything, slightly harder on Claude models; four-paper bootstrap $[-0.032, +0.003]$); over all nodes the four within-paper contrasts are within ± 0.006. These statements are limited to the four re-judged papers and ten models. The same node answer is scored against every leaf beneath it; across those leaves its scores have a mean standard deviation of 0.12 (450 internal-node answers: 45 internal nodes on these four papers, ten models), which mixes judge noise with genuine differences in how much of each leaf the answer covers.

Table 2: **Judge swap by model.** Mean score under each judge over the same (lineage, node) pairs, their difference, and the per-pair Pearson correlation.

Model	n	Sonnet judge	Sol judge	diff	Pearson
Opus 5.5	361	0.789	0.841	-0.052	0.92
GPT-6 Astra	361	0.738	0.778	-0.040	0.95
GPT-5.6 Sol	361	0.732	0.773	-0.042	0.93
Sonnet 5	361	0.710	0.744	-0.034	0.94
GPT-5.6 Terra	361	0.722	0.775	-0.053	0.92
GPT-5.6 Luna	361	0.702	0.756	-0.054	0.94
GPT-5 mini	361	0.613	0.634	-0.021	0.93
GPT-5	361	0.637	0.670	-0.032	0.92
Haiku 4.5	361	0.446	0.472	-0.026	0.94
GPT-5 nano	361	0.279	0.297	-0.018	0.94

C.2 Disputed root problems

On three papers, GPT-6 Astra argued that the stated main result is false and gave a counterexample instead of a proof, which the rubric scores near zero. An audit found that Astra was right on two of them, because the builder had dropped a hypothesis or changed a convention when stating the root;

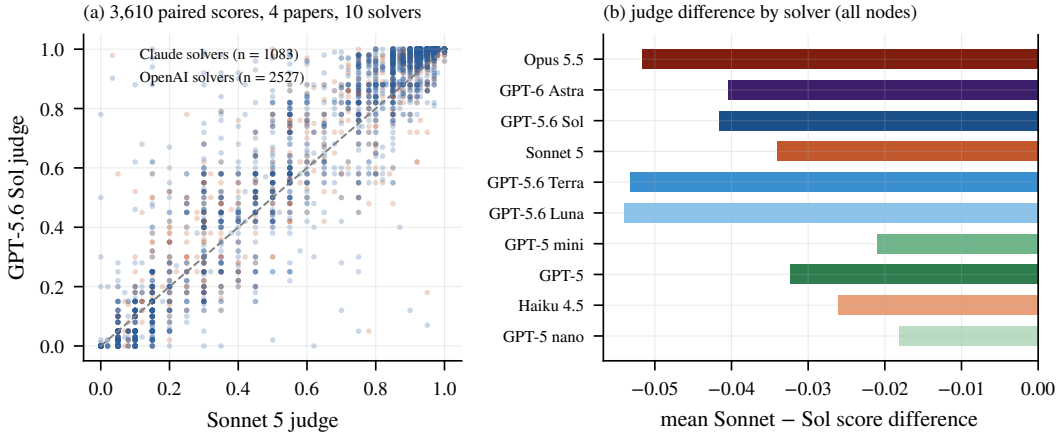

Figure 6: **Sonnet 5 vs. GPT-5.6 Sol judge.** (a) 3,610 paired (lineage, node) scores on four papers, ten models, coloured by model family; (b) mean Sonnet − Sol difference by model over all nodes.

on the third it rejects an exact reading of a step the paper uses as an approximation. We keep the original scores and report the audit, since automatically built trees will contain some misstated problems. Figure 7 compares the two strongest models paper by paper. On three papers GPT-6 Astra's root answer argues that the stated result is false and constructs a counterexample, and the judge, finding almost none of the paper's derivation, scores it 0.08. A keyword search of the judge's explanations flags these three cases and no other among the 200 (paper, model) pairs. We checked the three objections against the root statements and the papers with LLM assistance; this was not an audit by domain experts. For 2609.21840, the paper's width-hierarchy corollary requires adaptive width at least two, while the root's final comparison omits that condition; complement-of-singletons hypergraphs have adaptive and fractional hypertree widths tending to one and disprove the unrestricted logarithmic factor (the objection is correct about the root as posed). For 2609.10859, the root uses exact vanishing of the vertical kernel where the paper uses vanishing modulo base forms; a regular example satisfies the root's compatibility equations but its invariant algebra is not preserved by the supplied vector field (correct about the root as posed). For 2609.04992, an admissible continuous-density example has the same second moment as the proposed fractional diffusion equation but a different fourth moment, so the exact reduction asserted by the root does not hold; the paper imposes the same closure as an ansatz, so this objection concerns an exact reading of a phenomenological step. The other models repaired the logarithmic factor, adopted the relative kernel, added a preservation assumption, or proceeded with the closure. A direct scan of all 200 root answers (rather than the judge rationales) also finds whole-task rejections of 2609.10859 by GPT-5.6 Terra and Luna on a different, degree-convention argument, and partial objections embedded in otherwise affirmative answers (Astra on four further papers, Opus 5.5 and Sol on one), so refutation is not exclusive to Astra. No score was changed, and the objections beyond the three audited cases were not verified.

Excluding the disputed papers. We also exclude the complete trees of the two clearly misstated roots (2609.21840 and 2609.10859), then all three disputed roots (adding 2609.04992). These subsets were chosen after observing objections and are diagnostic sensitivities, not new benchmark definitions; the remaining problems have not been certified correct. The Opus–Astra whole-paper difference falls from $+0.110$ ($[+0.019, +0.218]$) to $+0.059$ ($[-0.013, +0.151]$) and $+0.024$ ($[-0.024, +0.068]$), respectively. On the latter subset, their atomic-step–whole-paper gaps remain 0.365 ($[0.270, 0.470]$) and 0.383 ($[0.291, 0.483]$). Thus the leading pair's separation is sensitive to the audited cases, whereas the endpoint gaps remain substantial. Resampling within each field on the full benchmark preserves the Opus–Astra mean difference, with interval $[+0.023, +0.214]$. All intervals are conditional on the retained scores and do not include reference, builder or judge uncertainty. The horizon under the same exclusions is in Appendix H.

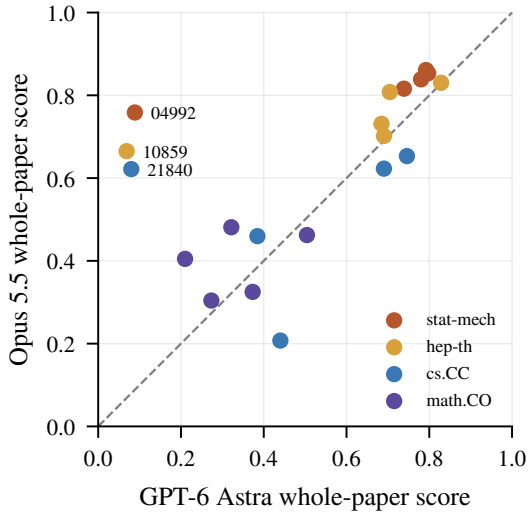

Figure 7: **Opus 5.5 against GPT-6 Astra, paper by paper** (whole-paper score; colour = field). The three labelled papers are those on which Astra refuted the stated result instead of deriving it; on two of them the root statement, not the paper, is at fault.

C.3 Harnesses and failures

Each lab's models ran in their own agent harness, so comparisons are between model and harness together; the two families also use the Python tool very differently (Appendix M). Failed answers are kept and scored as they are, which matters mainly for GPT-5 (Appendix M).

D A worked example: one tree end to end

This appendix shows one complete tree, built from arXiv:2609.15071, *The Komlós conjecture for complex discrepancy* (`math.CO`), which is also one of the four judge-swap papers. The tree has 28 nodes and 19 leaves. Table 3 lists every node with its size (leaves beneath), its importance weight and five models' weighted node scores $s(n)$; scores of at least 0.7 are bold. Every strong model reaches 0.7 on every leaf and most small nodes; no model reaches 0.7 on the 19-leaf root, the paper's main theorem.

Root problem (verbatim): the paper's main result, as posed to every model

Setting. For integers $m, n \geq 1$, let $\mathcal{A}(m,n)$ be the set of real matrices $A \in \mathbb{R}^{m\times n}$ all of whose columns have Euclidean norm at most 1 (real Komlós matrices), and let $\mathcal{C}(m,n)$ be the set of complex matrices $C \in \mathbb{C}^{m\times n}$ all of whose columns have Euclidean (ℓ_2) norm at most 1 (complex Komlós matrices). Write $\mathbb{T}^n = \{(e^{i\phi_1}, \ldots, e^{i\phi_n}) : \phi_j \in \mathbb{R}\}$ for the n-torus of unit-modulus complex vectors, and for $w \in \mathbb{C}^m$ let $\|w\|_\infty = \max_i |w_i|$ (largest modulus of an entry). Define:

- the classical (Boolean) discrepancy $\operatorname{disc}(A) = \min_{\varepsilon\in\{\pm1\}^n} \|A\varepsilon\|_\infty$;
- the complex discrepancy of $C \in \mathbb{C}^{m\times n}$: $\operatorname{cdisc}(C) = \min_{z\in\mathbb{T}^n} \|Cz\|_\infty$;
- the rank-2 vector discrepancy of $A \in \mathbb{R}^{m\times n}$: $\operatorname{Vdisc}_2(A) = \min_{u_1,\ldots,u_n\in\mathbb{S}^1} \max_{i\in[m]} \left|\sum_{j=1}^n A_{ij}u_j\right|$, where $\mathbb{S}^1 \subset \mathbb{R}^2$ is the unit circle and $|\cdot|$ is the Euclidean norm in $\mathbb{R}^2$.

The Komlós conjecture asserts that $\sup_{m,n} \max_{A\in\mathcal{A}(m,n)} \operatorname{disc}(A) < \infty$; it is open (the best known upper bounds grow with the dimension, e.g. $O(\sqrt{\log n})$ by Banaszczyk). Many relaxations have been studied, e.g. Gaussian discrepancy $\operatorname{Gdisc}(A) = \min_\sigma \mathbb{E}\|AG\|_\infty$ with $G \sim N(0,\sigma)$ and σ ranging over $n \times n$ positive semidefinite matrices with unit diagonal, which is known to satisfy $\operatorname{Gdisc}(A) \leq \sqrt{2}\operatorname{Vdisc}_2(A) \leq \sqrt{2}\operatorname{disc}(A)$ for real A. The complex Komlós problem asks whether $\max_{C\in\mathcal{C}(m,n)} \operatorname{cdisc}(C) = O(1)$ uniformly in m, n.

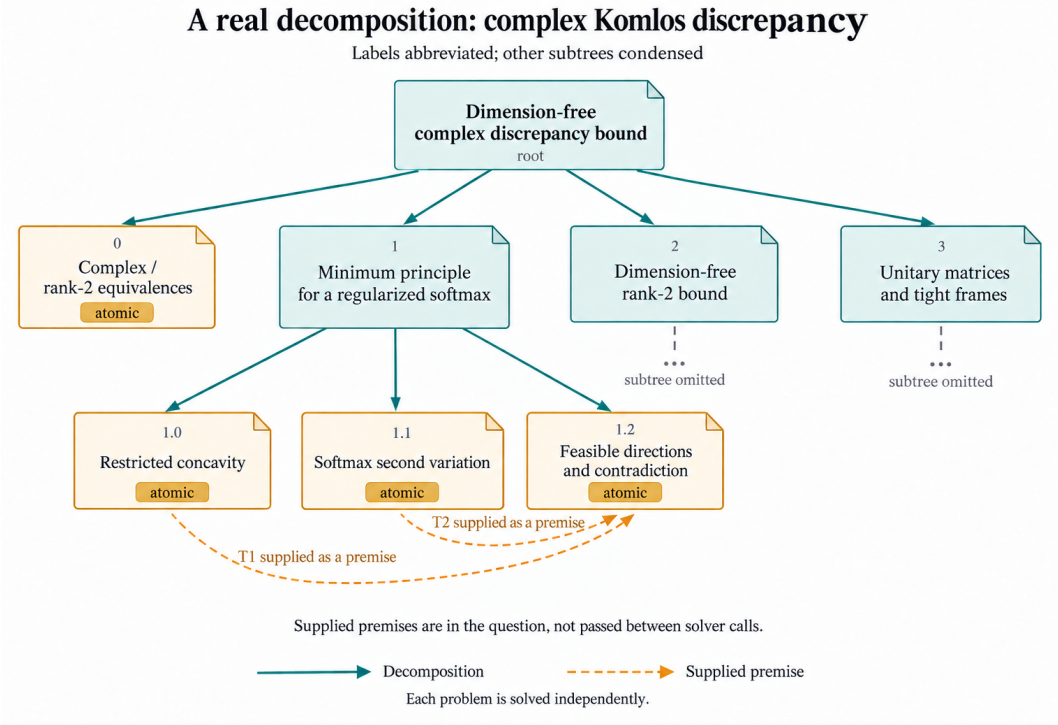

Figure 8: **What decomposition supplies to a model: a real tree excerpt.** The complex Komlós discrepancy tree, with node names abbreviated and only the minimum-principle branch expanded (full tree: Appendix D). Solid edges are the builder's decomposition; atomic tags mark terminal problems. Dashed edges show supplied premises: the statement of node `1.2` explicitly permits the restricted-concavity result from `1.0` (T1) and the softmax second-variation result from `1.1` (T2) without proof. The premises are part of node 1.2's question, not answers passed between model calls. These results are not supplied to the parent model; following the reference route requires establishing them, illustrating why the composition gap also reflects a change in the information provided. Colors encode node type, not measured performance.

Table 3: **The Komlós tree.** Node names are the builder's; sizes are leaves beneath the node; bold: $s(n) \geq 0.7$.

Node (indentation = depth)	Leaves	Weight	nano	GPT-5	Sonnet 5	Astra	Opus 5.5
Dimension-free bound for complex (rank-2) Komls discrepancy	19	1.00	0.22	0.26	0.29	0.27	0.30
Complex vs. rank-2 discrepancy: equivalences and consequences of a rank-2 bound	1	0.10	0.35	**0.98**	**0.97**	**0.98**	**0.98**
Abstract minimum principle for a regularized softmax under constraints	3	0.30	0.28	0.36	**0.87**	**0.86**	**0.87**
Restricted concavity: a direction where a block-trace-dominated PSD form is do…	1	0.35	0.35	**0.95**	**0.97**	**1.00**	**0.97**
Second variation of the softmax along directions equalizing directional deriva…	1	0.25	**0.95**	**0.95**	**1.00**	**0.98**	**0.97**
Constraint counting, feasible directions and the contradiction with local mini…	1	0.40	**0.93**	**0.95**	**0.97**	**0.97**	**0.98**
Dimension-free rank-2 Komls bound with constant 46 per coordinate	10	0.45	0.22	0.21	0.16	0.37	0.47
Sign, bound and monotonicity properties of the energy-regularized row potential	2	0.15	0.54	0.50	**0.95**	**0.96**	**0.95**
Negativity and sub-unit weight at zero discrepancy; discrepancy bound under th…	1	0.45	**0.92**	**0.97**	**0.97**	**1.00**	**1.00**
Monotonicity of the regularized potential and its exponential weight in the ac…	1	0.55	0.00	**0.95**	**1.00**	**1.00**	**1.00**
Contraction lemma: freezing columns preserves admissibility and does not incre…	2	0.25	0.40	**0.91**	**0.95**	**0.91**	**0.94**
Row-by-row case analysis: admissibility transfers and row weights do not incre…	1	0.70	0.00	**0.97**	**1.00**	**0.97**	**0.97**
From pointwise row-weight comparison to comparison of the softmax objective	1	0.30	0.00	**0.98**	**1.00**	**0.97**	**1.00**
Verifying the minimum-principle hypotheses for the admissible-set minimizer	3	0.35	0.47	**0.90**	**0.93**	**0.92**	**0.93**
Gradient and Hessian of the medium-row barrier z and weight gamma; exact quadr…	1	0.35	**0.97**	**0.98**	**0.97**	**0.98**	**1.00**
Averaged Hessian of the row potentials splits as K - D with block-trace bound	1	0.35	**0.88**	**0.97**	**0.98**	**1.00**	**0.97**
Constraint identification, counting hypothesis (A1) with mu = 4, and assemblin…	1	0.30	0.45	**0.95**	**0.97**	**0.96**	**0.95**
Global minimization over active sets and choice of constants yielding the boun…	3	0.25	0.01	0.40	**0.87**	**0.91**	**0.90**
Existence of a global minimizer over all active sets and the value bound at th…	1	0.35	**0.90**	**0.97**	**0.98**	**0.97**	**1.00**
The tie-broken global minimizer has empty active set	1	0.40	0.00	**0.97**	**0.97**	**0.98**	**0.95**
Choosing explicit constants compatible with C = 46 and concluding the discrepa…	1	0.25	**0.95**	**0.95**	**0.95**	**0.97**	**0.97**
Complex discrepancy of unitary matrices and tight frames	5	0.15	0.63	0.69	**0.92**	**0.89**	**0.91**
Singular-value lower bound on complex discrepancy	1	0.20	**0.95**	**0.97**	**1.00**	**0.97**	**1.00**
Upper bound on complex discrepancy by the infinity-norm of $(CC^*)^{1/2}$	3	0.55	0.55	**0.82**	**0.93**	**0.95**	**0.89**
Biunimodular vectors of unitary matrices	1	0.30	**0.95**	**0.95**	**1.00**	**1.00**	**0.98**
Factoring C through $(CC^*)^{1/2}$ and a co-isometry	1	0.40	**0.95**	**0.97**	**0.97**	**1.00**	**0.97**
Concluding the infinity-norm bound from the factorization	1	0.30	**0.97**	**0.97**	**1.00**	**0.98**	**0.98**
Tight frames and the exact complex discrepancy of unitary matrices	1	0.25	**0.97**	**0.98**	**0.97**	**1.00**	**1.00**

Task. Prove the following.

1. (Equivalences.) For real A, identifying $z_j = e^{i\phi_j}$ with $u_j = (\cos\phi_j, \sin\phi_j)$, show $\operatorname{cdisc}(A) = \operatorname{Vdisc}_2(A)$. With $K_2(m,n) := \max_{A\in\mathcal{A}(m,n)} \operatorname{Vdisc}_2(A)$, show
$$K_2(m,n) \leq \max_{C\in\mathcal{C}(m,n)} \operatorname{cdisc}(C) \leq 2K_2(2m,n).$$
2. (Main theorem: rank-2 Komlós.) Prove that there is an absolute, explicit constant bounding the rank-2 vector discrepancy of every real Komlós matrix, independently of m and n; specifically, for every m, n and every $A \in \mathcal{A}(m,n)$ there exist unit vectors $x_1, \ldots, x_n \in \mathbb{R}^2$ with
$$\max_{i\in[m]} \Big| \sum_{j=1}^{n} A_{ij} x_j \Big| \leq 46\sqrt{2},$$
and in fact with every coordinate of each row sum $\sum_j A_{ij}x_j \in \mathbb{R}^2$ bounded in absolute value by 46. Deduce that $\operatorname{cdisc}(C) \leq 92\sqrt{2}$ for every complex Komlós matrix $C \in \mathcal{C}(m,n)$ (thus resolving the complex Komlós problem), and that the Komlós conjecture holds for Gaussian discrepancy: $\operatorname{Gdisc}(A) \leq 92$ for all real Komlós matrices.
3. (Unitary matrices.) Show that for every $C \in \mathbb{C}^{m\times m}$, $\sigma_{\min}(C) \leq \operatorname{cdisc}(C)$ (where $\sigma_{\min}$ is the smallest singular value), and $\sigma_{\min}(C) \leq 1$ for $C \in \mathcal{C}(m,m)$. Show more generally that for $C \in \mathbb{C}^{m\times n}$ with $m \leq n$, $\operatorname{cdisc}(C) \leq \|(CC^*)^{1/2}\|_{\infty\to\infty}$, where $\|M\|_{\infty\to\infty} = \max_i \sum_j |M_{ij}|$ and $(CC^*)^{1/2}$ is the positive semidefinite square root; conclude that $\operatorname{cdisc}(C) \leq \sqrt{\lambda}$ whenever $CC^* = \lambda I_m$, and hence

$\mathrm{cdisc}(U) = 1$ exactly for every $m \times m$ unitary matrix U. You may use the following fact (Idel–Wolf): for every unitary $U \in \mathbb{C}^{m\times m}$ there exist diagonal unitary matrices D_1, D_2 such that $W = D_1 U D_2$ has all row sums and all column sums equal to 1.

The leaf below lies on the lineage root → `1` (the abstract minimum principle behind the main bound) → `1.0`.

Leaf problem (verbatim): node `1.0`, "Restricted concavity: a direction where a block-trace-dominated PSD form is dominated by a block-scalar form"

Let $s \geq 1$ and identify $\mathbb{R}^{2s}$ with $(\mathbb{R}^2)^s$, writing $v = (v_1, \ldots, v_s)$ with $v_j \in \mathbb{R}^2$, with the standard inner product $(\cdot,\cdot)$. For a matrix $M \in \mathbb{R}^{2s\times 2s}$ let $M_{jj} \in \mathbb{R}^{2\times 2}$ denote its j-th 2×2 principal diagonal block. Let $d_1, \ldots, d_s \geq 0$ and let $D = \mathrm{diag}_j(d_j I_2)$ be the block-diagonal matrix whose j-th diagonal block is $d_j I_2$ (so $D \succeq 0$). Let $K \in \mathbb{R}^{2s\times 2s}$ be symmetric positive semidefinite, and let $\rho > 0$ be a constant such that

$$\mathrm{tr}\, K_{jj} \leq \rho\, d_j \qquad \text{for every } j = 1, \ldots, s.$$

Prove that for every linear subspace $W \subseteq \mathbb{R}^{2s}$ with $\dim W > \rho s$ there exists a nonzero $v \in W$ with

$$(v, Kv) \leq (v, Dv).$$

Its reference solution, extracted from the paper (first lines, verbatim)

Solution: restricted concavity (Lemma 21 of the paper)

Claim. Let $D = \mathrm{diag}_j(d_j I_2) \succeq 0$, $K \succeq 0$ in $\mathbb{R}^{2s\times 2s}$ with $\mathrm{tr}\, K_{jj} \leq \rho d_j$ for all j. Then every subspace W with $\dim W > \rho s$ contains a nonzero v with $(v, Kv) \leq (v, Dv)$.

The proof below is the paper's argument, EXTRACTED, with a few small steps written out in full.

Step 1: $\ker D \subseteq \ker K$ Fix j with $d_j = 0$. Then $\mathrm{tr}\, K_{jj} \leq 0$. The block K_{jj} is a principal submatrix of the PSD matrix K, so it is PSD and its trace is ≥ 0. Hence $\mathrm{tr}\, K_{jj} = 0$. A PSD matrix with zero trace has all eigenvalues zero, so $K_{jj} = 0$. Now use the standard fact that if a PSD matrix has a zero diagonal entry $K_{aa} = 0$, then its whole row and column a vanish. This follows from the 2×2 minors: $K_{ab}^2 \leq K_{aa} K_{bb} = 0$. So the entire ...

Lineage judging. For this leaf the judge scores every answer on the lineage against the same reference solution. Opus 5.5's answer to the leaf itself scores 0.97, its answer to node `1` 0.88, and its answer to the root, which had to find this step inside a proof of the whole theorem, 0.10. The pattern holds across the tree: against the 6 leaves under nodes `0` and `3` (the equivalences and the unitary case, parts 1 and 3 of the root task) Opus 5.5's root answer scores 0.80–0.95, against the 13 leaves of the main theorem (nodes `1` and `2`) it scores 0.05–0.40, and posed alone the same 13 leaves score 0.95–1.00. GPT-6 Astra shows the same split. Posed the whole paper, both cover the subsidiary parts but little of the main proof; posed the steps one by one, both solve them almost perfectly. The judge's rationales (verbatim, truncated):

Judge on Opus 5.5's root answer, against leaf `1.0`: score 0.10

"The root answer addresses the complex/rank-2 Komlos reductions and unitary matrix results (equivalence of cdisc and Vdisc_2, the 2K_2 bound, Gaussian discrepancy consequence, unitary discsic=1), and explicitly states it could not reconstruct a proof of the core estimate. It contains no discussion of the linear-algebra restricted concavity lemma (matrices K, D, trace bound tr K_jj ¡= rho d_j, subspace dimension argument) that the leaf problem asks for. None of the key elements (ker D subset ker K, pseudoinverse/rescaling argument, eigenvalue counting, dimension count) are present, so it does ..."

Judge on Opus 5.5's answer to leaf `1.0` itself: score 0.97

"This is the leaf node itself and gives essentially the same proof as the ground-truth solution: it establishes ker D subset ker K via the same PSD zero-trace/zero-row argument, splits into the case W meets ker D

(trivial equality) versus not, reduces to the subblock U with all d_j¿0, forms A = D{}{-1/2}KD{}{-1/2} with trace ¡= rho*s, bounds the number of eigenvalues exceeding 1, and uses a . . . ”

E THE ALGORITHMS, STEP BY STEP

Algorithm 1 gives the whole pipeline; Figure 9 runs the decomposition-curve computation on the Komlós tree of Appendix D with Opus 5.5's answers. Cuts are indexed by merge level, so F_0 (all steps) sits at $x = 1$ and F_3 (the whole paper) at $x = 0$.

Algorithm 1 RESEARCHTREE: build, solve blind, judge lineages, read out

Require: paper P; builder B; model M; judge J; success threshold τ ($= 0.7$; only for the thresholded variant H^{τ}_{50})

Build. $n_0 \leftarrow B$(state the main result of P as one self-contained problem)
queue $\leftarrow [n_0]$
while queue not empty **do**
  $n \leftarrow$ pop(queue)
  if B marks n atomic, or depth$(n) = 6$, or the tree has 400 nodes **then**
    $y_n \leftarrow B$(extract the solution of n from P, marking extracted and inferred steps) {leaf reference}
  else
    children $c_1..c_k \leftarrow B$(split n into $2 \le k \le 4$ self-contained sub-problems; a sibling's result enters only as an explicit ASSUME clause)
    importance weights $w(n \to c_i) \ge 0$ with $\sum_i w(n \to c_i) = 1$; append $c_1..c_k$ to queue
  end if
end while
Solve blind. for all nodes n: $a(n) \leftarrow M$(statement of n only; Python sandbox; no paper, no tree)
Judge lineages. for all leaves L with lineage $n_0, \ldots, n_d = L$:
  $\mathrm{Score}(n_k, L) \leftarrow J\big(a(n_k), \mathrm{problem}(L), y_L\big)$ for $k = 0, \ldots, d$ {one call, each answer scored separately}
Node scores. $s(n) \leftarrow \sum_{L \in \mathcal{L}(n)} w(n \to L)\,\mathrm{Score}(n, L) \,/\, \sum_{L \in \mathcal{L}(n)} w(n \to L)$
Decomposition curve. for ten random orders: $F \leftarrow$ all leaves; record $(|F|, R(F))$ with $R(F) = \sum_{n \in F} w(n_0 \to n)\, s(n)$
  while $F \neq \{n_0\}$: pick, among parents whose children are all in F, one at the greatest depth (random tie-break); replace its children in F by it; record $(|F|, R(F))$
  average the curves over orders on $x = (|F| - 1)/(L - 1)$; decomposition deficit $D \leftarrow 1-$ mean of R over x; share of $R_{\mathrm{full}} - R_{\mathrm{root}}$ recovered at each depth cut
Horizon. fit $\mathbb{E}[s(n)] = \sigma(\alpha + \beta \log_2 |\mathcal{L}(n)|)$ over all nodes of all papers (fractional logistic); $H_{50} \leftarrow 2^{-\alpha/\beta}$
  uncertainty: resample papers with replacement (1,000 draws), refit, report percentile intervals; thresholded variant $\Pr[s(n) \ge 0.7]$ with 2,000 draws

F PROMPTS

The builder, model and judge instructions below are reproduced verbatim from the workflow code that ran the experiments (file paths in the released code). The model additionally receives the path of its node's `problem.md`; the judge receives the run directory, the lineage's leaf problem and reference solution, and the ancestors' answers.

Builder: root problem

```
You are analyzing a theoretical research paper to identify its central
research question or main result, as the root of a ResearchTree problem
tree.  You will be given RUN_DIR. Read `RUN_DIR/paper.md` in full.
FIGURES: the markdown links each figure as `![Figure N
...](<id>_figures/<file>.png)` right above its verbatim caption; the
PNG files live in `RUN_DIR/figures/`.  When a statement, proof or
derivation depends on what a figure shows (a diagram of the
construction, a phase diagram, a graph gadget, a plotted comparison),
```

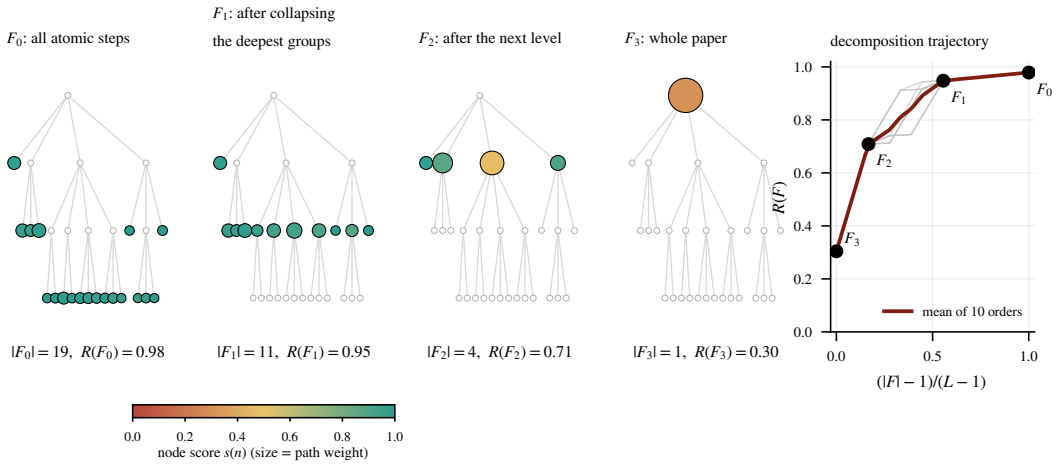

Figure 9: **How a decomposition curve is computed**, on the 19-leaf Komlós tree (Appendix D) with Opus 5.5's answers. Cuts are indexed by merge level, so F_0 (all steps) sits at $x = 1$ and F_3 (the whole paper) at $x = 0$. Each cut F_k covers every leaf exactly once; its nodes are filled, coloured by their score $s(n)$ and sized by their path weight $w(n_0 \to n)$, and $R(F_k)$ is their weighted sum. F_0 poses the paper as its 19 atomic steps ($R = 0.98$); collapsing the deepest sibling groups leaves 11 sub-problems (0.95); at the next level the main-proof branch drops to about 0.5 and $R = 0.71$; posed whole, the paper scores 0.30. Right: the ten random orders (grey) and their mean (the decomposition curve), with the four cuts marked; the level boundaries are the same in every order.

```
open that image with Read and use its content; do not guess from the
caption alone.

Formulate the paper's main contribution as a COMPLETE, SELF-CONTAINED
problem statement.

1.  FULLY SELF-CONTAINED: include ALL necessary definitions, notation,
and context.  A reader with NO access to the paper must fully
understand the problem.  NO references to "Theorem X", "Equation Y",
"Section Z", or "the paper".  No appeals to "the known/standard X" for
paper-specific objects:  write them out.  Every symbol used (including
Lax operators, matrix units, spectral parameters) must be defined in
the statement.  2.  CAPTURE THE ENTIRE PAPER: solving this problem
should essentially reproduce the paper's main contribution.  Provide
sufficient background; several paragraphs are expected.  3.  DO NOT
HINT AT THE SOLUTION: state what needs to be shown, derived, or proven,
not how the paper does it.  4.  STYLE-AGNOSTIC: main theorem to prove
(math), main result to derive or establish (physics), main algorithmic
or complexity result (TCS).

Write `RUN_DIR/nodes/root/node.json` with exactly these keys:  {"path":
"", "depth":  0, "name":  "<5-10 word title>", "problem_statement":
"<complete statement, LaTeX with $...$ allowed>", "weight":  1.0,
"link_description":  "", "atomic":  false}

Then reply with one line:  the node name and the character count of the
problem statement.
```

Builder: decomposition

```
You are decomposing a research problem into sub-problems for a
ResearchTree problem tree.  You will be given RUN_DIR, PARENT_SLUG,
PARENT_DEPTH, MAX_DEPTH, and optionally BUDGET guidance.  Read
`RUN_DIR/paper.md` (reference for finding the logical sub-problems) and
```

```
`RUN_DIR/nodes/PARENT_SLUG/node.json` (the parent problem).  FIGURES:
the markdown links each figure as `![Figure N
...](<id>_figures/<file>.png)` right above its verbatim caption; the
PNG files live in `RUN_DIR/figures/`.  When a statement, proof or
derivation depends on what a figure shows (a diagram of the
construction, a phase diagram, a graph gadget, a plotted comparison),
open that image with Read and use its content; do not guess from the
caption alone.

YOUR TASK: identify the logical sub-problems needed to solve the PARENT
PROBLEM. The paper contains or references the steps and components
needed; extract them as sub-problems.

CRITICAL REQUIREMENTS 1.  DECOMPOSE ONLY THE PARENT PROBLEM. Focus
exclusively on what is needed to solve the parent.  Do not re-extract
the whole paper.  The parent's siblings (listed when present) are being
decomposed separately, so it is best not to overlap with them; do not
mention them in a statement.  2.  LOGICAL STRUCTURE, not paper section
order.  Decompose by logical dependencies; group steps sharing one
technique or that are special cases of each other.  If result B is a
corollary of A, they belong together.  3.  PROBLEM STATEMENTS ARE
QUESTIONS, NOT ANSWERS. State WHAT must be shown, derived, or computed.
Do not leak the method or the result.  GOOD: "Show that the integral
converges for all x > 0".  BAD: "...using dominated convergence" (leaks
method).  BAD: "Show that the integral equals pi/2" (leaks answer).  4.
SELF-CONTAINED (most important).  Each sub-problem must be
understandable in COMPLETE ISOLATION by a reader who sees only that
sub-problem:  not the parent, not the siblings, not the paper.  Include
ALL definitions, equations, and notation it needs; repeat shared
equations in every sub-problem that needs them.  NEVER reference other
sub-problems.  If B needs a result from A, state that result in B as an
explicit assumption ("Assume that ...  holds.  Under this assumption,
...").  FORBIDDEN phrases:  "the previous subproblem", "earlier
subproblem", "as before", "as shown above", "the above", "as
mentioned", "from the paper", "in the previous part", "using the same
notation as", "with the same definitions", "as in the parent problem",
"continuing from", "from part (a)", "the result established earlier",
"we showed that", "it was proven that".  ALSO FORBIDDEN: appeals to
unstated knowledge such as "the known/standard/usual X", "the
well-known description", "the familiar construction" when X is
paper-specific; either write X out or drop it.  Every symbol that
appears (matrix units, spectral parameters, Lax operators, generating
functions, special functions) must be defined inside this statement;
re-read your statement and check each symbol before writing the file.
5.  BRANCHING: typically 2-3 sub-problems, 4 when the parent genuinely
splits four ways; use your judgment about what the logical structure
needs, never pad to a fixed number.  Each must be a meaningful chunk,
not a tiny step.  6.  IMPORTANCE WEIGHTS in [0,1] reflecting importance
for solving the parent; weights must sum to 1.0.  7.  ATOMICITY: use
your judgment.  Mark atomic=true when the sub-problem is a
self-contained, exam-style question that a well-prepared graduate
student could answer as one exam question:  a single computation, one
inequality or identity, one limit, one case check, one short argument.
Its complete solution is one argument of about half a page to one page,
with no sub-lemmas.  Mark atomic=false when solving it needs several
distinct steps each needing real work; a whole lemma with its proof, or
a construction plus its analysis, is a chunk to decompose further, not
an exam question.  Atomic sub-problems become leaves, so leaves may sit
at different depths.  If PARENT_DEPTH + 1 >= MAX_DEPTH, every child
must be marked atomic=true and stated so that it is solvable as one
exam question.

Write one file per child, `RUN_DIR/nodes/PARENT_SLUG_i/node.json` for i
= 0, 1, 2, ...  (slug = parent slug + "_" + i; if the parent slug is
```

```
"root", the child slug is just "i"), each with exactly these keys:
{"path": "<parent path>/<i>" (or "<i>" if parent is root), "depth":
PARENT_DEPTH + 1, "name": "<brief title>", "problem_statement":
"<complete self-contained statement>", "weight": <float>,
"link_description": "<how this sub-problem contributes to solving the
parent>", "atomic": <true|false>}

Finish by returning, as your final output, the structured result
requested (list of children with slug, path, depth, weight, atomic,
name). Do not write anything else.
```

Builder: leaf ground-truth solutions

```
You are extracting ground-truth solutions for leaf sub-problems of a
ResearchTree problem tree. You will be given RUN_DIR and a list of
LEAF_SLUGS. Read `RUN_DIR/paper.md` once, then for each slug read
`RUN_DIR/nodes/<slug>/node.json` and extract the COMPLETE solution to
its problem_statement from the paper. FIGURES: the markdown links each
figure as `![Figure N ...](<id>_figures/<file>.png)` right above its
verbatim caption; the PNG files live in `RUN_DIR/figures/`. When a
statement, proof or derivation depends on what a figure shows (a
diagram of the construction, a phase diagram, a graph gadget, a plotted
comparison), open that image with Read and use its content; do not
guess from the caption alone.

Each solution must include: all relevant reasoning and key steps; the
mathematical derivations; important intermediate results; the final
answer or result; crucial assumptions and conditions. It should be
complete but concise (a model answer), self-contained (understandable
without the paper), and directly address the stated problem. Use LaTeX
with $...$ and $$...$$.

If the paper does not contain complete information (a derivation left
as an exercise, a "straightforward" skipped step), fill the gap and
solve it, but clearly mark what is EXTRACTED from the paper and what
you INFERRED. If the paper genuinely does not contain and does not
imply the required content, say so explicitly in a first line "STATUS:
paper does not contain this result" and still give the best
reconstruction you can.

Write each solution to `RUN_DIR/nodes/<slug>/solution.md`. A solution
shorter than 500 characters is almost never complete; expand it. Do
not end a solution with a bare separator line. When done, return the
structured result requested (per slug: characters written and status
in {extracted, partly_inferred, not_in_paper}). Do not write anything
else.
```

Solver

```
You are solving a research problem from a theoretical paper. You will
be given a path to a file `problem.md`. Read it, then write your
complete answer to `answer.md` in the same directory. Do not read any
other file, do not search, do not open the paper; the whole point is a
blind evaluation on the problem statement alone.

Code interpreter: you may run Python (sympy, numpy, scipy, mpmath
available) ONLY through the sandboxed wrapper, from inside
nodes/<slug>/work/: `bash <repo>/${PYRUN} script.py` or `... pyrun.sh
-c "code"`. It has no network and a 300 s limit. Use it to check
algebra, series, integrals, and numerics. Never call curl, wget, pip,
git, ssh, or python directly, and never touch any `_judge/` or `data/`
path. All tool calls are audited; a violation voids the answer.
```

```
The problem is part of a hierarchical decomposition of a larger
research question.  Your task is to:  1.  Carefully analyze what the
problem is asking 2.  Explore all relevant directions thoroughly 3.
Provide a detailed, complete answer that covers all aspects

Since this comes from a research paper, you should:  - Define all terms
and notation you use - Show your reasoning step by step with full
mathematical derivations - Write out all proofs in complete detail; do
not skip steps or say "it follows that" - Include all intermediate
calculations and algebraic manipulations - Cover edge cases and special
considerations - Be comprehensive; your answer should contain all the
details needed to fully address the problem

Do not be brief.  A thorough, detailed answer with complete
mathematical proofs and derivations is required.  Treat this as writing
a complete solution for a research-level exam where every step must be
justified.  Write LaTeX math with $...$ and $$...$$.  When finished,
reply with one line:  the answer file path and its character count.
```

Judge

```
You are evaluating whether a model's answer to a research problem
contains the key elements needed to derive a specific leaf solution.
You will be given a run directory and a lineage id.  Read
`_judge/paper.md` (context for terminology and notation only),
`_judge/lineages/<id>/leaf.json` (leaf problem, ground-truth leaf
solution, and the list of ancestor node slugs from root to leaf), and
for each ancestor slug read `nodes/<slug>/answer.md`.

Score EACH ancestor answer INDEPENDENTLY against the same leaf
solution.  Do not let one answer's score influence another; do not
grade on a curve.  An ancestor at depth 0 (root) answered the whole
paper's problem, so judge only whether its content covers what the leaf
solution requires.

Scoring rubric (0.0-1.0):  - 0.9-1.0:  All key elements present:
definitions, main steps, and results align with the leaf solution -
0.7-0.9:  Right direction, good foundation, most details present,
missing a few specifics - 0.4-0.6:  Core insight present, heading
toward the solution, but significant details missing - 0.1-0.3:
Tangentially related concepts, not really building toward this solution
- 0.0-0.1:  Unrelated or contradicts the solution

Focus on comprehensiveness (does the answer address what the leaf
problem requires), details (key steps, definitions, results present),
and direction (is it building toward this solution).  If the
ground-truth solution says the paper does not contain the answer, give
credit only for correct derivations, and 0 for invented content that
contradicts it.

Write `_judge/lineages/<id>/judgments_<tag>.json` exactly in this form:
{"lineage_id":  "<id>", "judgments":  [{"slug":  "<slug>", "depth":
<int>, "score":  <float>, "reasoning":  "<2-5 sentences>"}, ...]}
Include one entry per ancestor slug.  Then reply with one line listing
the scores from root to leaf.
```

G TREES AND PER-PAPER RESULTS

Paper selection. We listed all 881 papers posted in September 2026 in the four categories, dropped very short and very long ones by page count, and screened the rest with an LLM on title and abstract for self-contained results that can be derived without experiments or data. Candidates were converted to text, and an LLM read each full paper and scored whether its main result can be checked by derivation alone and how much work it represents. Between 23 (`cs.CC`) and 74 (`math.CO`)

papers per field qualified; this study uses five per field. The same filter can be rerun on any later month of postings.

Trees. Claude Opus 5.5 built 18 of the trees. The other two (2609.15133 and 2609.10859) come from an earlier builder run, mostly by Claude Opus 5, and were kept because their evaluations were already complete. Every model answered the same tree for every paper. Weighting leaves by importance raises every whole-paper score by about 0.04 relative to an unweighted average and leaves the ranking of models unchanged. Table 4 lists the trees; Table 5 and Table 6 give the whole-paper and step scores of every model on every paper. Individual papers mostly agree with the overall ranking: a paper's ranking of the models correlates with that of the other 19 papers at 0.81 on average (Spearman, $[0.71, 0.90]$).

Table 4: **The 20 trees.** One lineage per leaf, so lineages equal leaves.

Category	Paper	Nodes	Leaves	Max depth	Lineages
stat-mech	2609.04668	54	38	6	38
stat-mech	2609.04992	39	26	4	26
stat-mech	2609.05247	51	36	4	36
stat-mech	2609.07496	54	38	5	38
stat-mech	2609.16821	40	28	6	28
cs.CC	2609.10233	29	20	4	20
cs.CC	2609.10289	25	17	3	17
cs.CC	2609.18020	57	36	4	36
cs.CC	2609.19623	36	24	4	24
cs.CC	2609.21840	21	15	3	15
hep-th	2609.06363	61	42	4	42
hep-th	2609.10272	43	30	4	30
hep-th	2609.10823	36	25	4	25
hep-th	2609.10859	186	130	6	130
hep-th	2609.11417	51	35	4	35
math.CO	2609.07285	64	43	5	43
math.CO	2609.15071	28	19	3	19
math.CO	2609.15133	87	63	5	63
math.CO	2609.15201	68	46	6	46
math.CO	2609.15270	49	33	4	33
Total	20 papers	1079	744		744

Table 5: **Whole-paper (root) score per paper and model.**

Cat.	Paper	Opus	Astra	Sol	Son.	Terra	Luna	mini	GPT-5	Haiku	nano
stat-mech	2609.04668	0.86	0.79	0.76	0.79	0.73	0.78	0.58	0.74	0.45	0.29
stat-mech	2609.04992	0.76	0.09	0.47	0.48	0.70	0.43	0.27	0.54	0.17	0.20
stat-mech	2609.05247	0.84	0.78	0.73	0.81	0.65	0.68	0.60	0.52	0.30	0.00
stat-mech	2609.07496	0.82	0.74	0.75	0.80	0.70	0.75	0.73	0.62	0.47	0.23
stat-mech	2609.16821	0.85	0.80	0.73	0.77	0.66	0.72	0.62	0.00	0.55	0.15
cs.CC	2609.10233	0.62	0.69	0.59	0.54	0.51	0.49	0.31	0.36	0.25	0.00
cs.CC	2609.10289	0.65	0.75	0.62	0.40	0.56	0.62	0.30	0.31	0.20	0.01
cs.CC	2609.18020	0.21	0.44	0.22	0.21	0.23	0.22	0.14	0.19	0.15	0.17
cs.CC	2609.19623	0.46	0.38	0.17	0.11	0.17	0.14	0.11	0.10	0.07	0.10
cs.CC	2609.21840	0.62	0.08	0.33	0.29	0.32	0.29	0.23	0.26	0.16	0.10
hep-th	2609.06363	0.70	0.69	0.50	0.49	0.48	0.46	0.36	0.40	0.33	0.31
hep-th	2609.10272	0.81	0.70	0.68	0.61	0.61	0.67	0.54	0.52	0.27	0.21
hep-th	2609.10823	0.83	0.83	0.76	0.73	0.74	0.71	0.58	0.69	0.34	0.00
hep-th	2609.10859	0.67	0.07	0.46	0.57	0.05	0.05	0.25	0.27	0.28	0.23
hep-th	2609.11417	0.73	0.68	0.60	0.61	0.51	0.58	0.44	0.48	0.34	0.00
math.CO	2609.07285	0.48	0.32	0.26	0.20	0.20	0.20	0.22	0.26	0.15	0.15
math.CO	2609.15071	0.30	0.27	0.31	0.29	0.28	0.30	0.23	0.26	0.21	0.22
math.CO	2609.15133	0.33	0.37	0.24	0.20	0.21	0.24	0.13	0.18	0.11	0.13
math.CO	2609.15201	0.46	0.50	0.41	0.25	0.25	0.28	0.24	0.22	0.19	0.00
math.CO	2609.15270	0.40	0.21	0.23	0.25	0.23	0.23	0.16	0.02	0.14	0.01
	mean	0.62	0.51	0.49	0.47	0.44	0.44	0.35	0.35	0.26	0.13

Table 6: **Atomic-step (leaf) score per paper and model.**

Cat.	Paper	Opus	Astra	Sol	Son.	Terra	Luna	mini	GPT-5	Haiku	nano
stat-mech	2609.04668	0.98	0.97	0.98	0.98	0.97	0.97	0.95	0.96	0.86	0.52
stat-mech	2609.04992	0.98	0.95	0.97	0.97	0.97	0.95	0.96	0.95	0.63	0.39
stat-mech	2609.05247	0.98	0.98	0.98	0.98	0.97	0.98	0.95	0.96	0.82	0.58
stat-mech	2609.07496	0.98	0.97	0.97	0.97	0.97	0.95	0.96	0.95	0.90	0.18
stat-mech	2609.16821	0.98	0.97	0.97	0.98	0.96	0.96	0.96	0.92	0.76	0.40
cs.CC	2609.10233	0.97	0.97	0.97	0.97	0.97	0.96	0.95	0.96	0.83	0.66
cs.CC	2609.10289	0.98	0.98	0.97	0.97	0.98	0.97	0.96	0.96	0.84	0.42
cs.CC	2609.18020	0.98	0.97	0.97	0.96	0.97	0.97	0.91	0.85	0.70	0.58
cs.CC	2609.19623	0.97	0.96	0.97	0.97	0.96	0.96	0.94	0.90	0.78	0.45
cs.CC	2609.21840	0.98	0.97	0.97	0.97	0.97	0.97	0.94	0.96	0.80	0.42
hep-th	2609.06363	0.97	0.97	0.97	0.97	0.96	0.97	0.92	0.87	0.77	0.64
hep-th	2609.10272	0.98	0.94	0.97	0.97	0.96	0.96	0.94	0.94	0.77	0.50
hep-th	2609.10823	0.96	0.97	0.97	0.95	0.97	0.95	0.92	0.96	0.74	0.67
hep-th	2609.10859	0.98	0.94	0.96	0.96	0.93	0.94	0.88	0.88	0.76	0.44
hep-th	2609.11417	0.98	0.96	0.96	0.95	0.96	0.93	0.83	0.86	0.69	0.18
math.CO	2609.07285	0.97	0.97	0.97	0.96	0.97	0.97	0.91	0.87	0.78	0.45
math.CO	2609.15071	0.98	0.98	0.98	0.98	0.98	0.97	0.95	0.96	0.84	0.61
math.CO	2609.15133	0.97	0.97	0.97	0.95	0.96	0.96	0.80	0.91	0.59	0.45
math.CO	2609.15201	0.97	0.98	0.97	0.97	0.97	0.97	0.94	0.89	0.75	0.47
math.CO	2609.15270	0.96	0.97	0.95	0.96	0.95	0.95	0.88	0.91	0.83	0.28
	mean	0.98	0.97	0.97	0.97	0.97	0.96	0.92	0.92	0.77	0.46

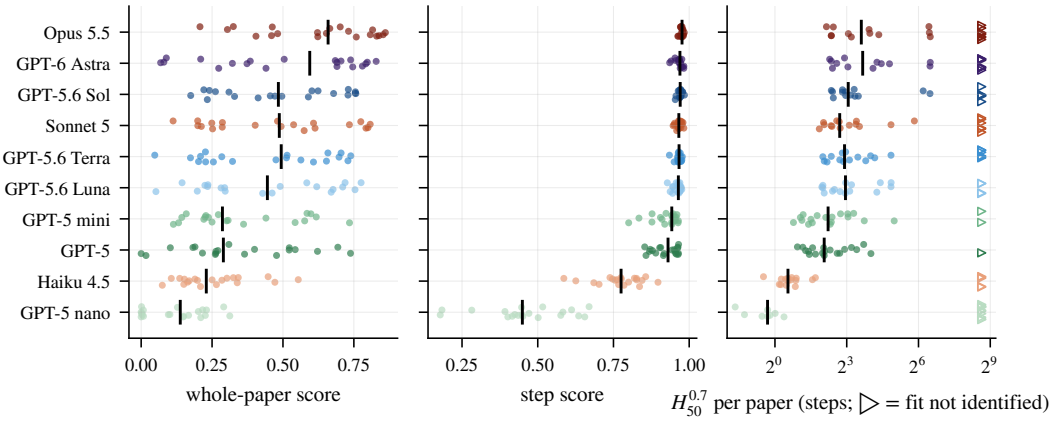

Figure 10: **Per-paper spread.** Weighted whole-paper score, weighted step score and per-paper thresholded horizon $H_{50}^{0.7}$ for every model; one dot per paper, black tick = median over papers; triangles mark papers whose per-paper fit is not identified (all successes or separation).

H Measurement 1: Research Horizon Details

H.1 The Research Horizon H_{50}

The main fit. The main text fits the fractional logistic $\mathbb{E}[s(n)] = \sigma(\alpha + \beta \log_2 |\mathcal{L}(n)|)$ to the node scores themselves (quasi-binomial likelihood, each node weighted equally, no clipping or threshold) and defines H_{50} as the size at which the fitted expected score is one half; intervals resample papers (1,000 draws; papers resampled within the field for the field fits). Table 7 gives every model's horizon and slope, overall and by field. The slopes of the eight models above Haiku 4.5 lie between -0.65 and -0.75 logits per doubling, which is why rescaling by the horizon collapses their curves (main-text Figure 3d; for the eight capable models, binned-mean RMS 0.048 around one logistic against 0.084 unrescaled, and 0.076 against 0.156 for model $\times$ field). For the four leading releases this horizon doubles every 7.9 months ($[6.8, 9.5]$); leaving out GPT-5 gives 3.1 months and leaving out any other release 7.6–9.1; the mid tier gives 13.6 months ($[11.3, 16.9]$). The eight capable models' slopes are similar but not identical; a clipped-logit regression with paper fixed effects resolves differences of about 0.1 logits per doubling (Appendix K), so the collapse is approximate.

Robustness of H_{50}. Removing the complete trees of the two misstated root problems, or of all three disputed papers (Appendix C.2), lowers Opus 5.5's horizon from 38.1 to 31.1 and 30.5 steps and changes every other capable model by at most 2.1 steps, so Opus 5.5 stays first and the ranking of the others is nearly unchanged; weighting papers equally gives 31.2 for Opus 5.5. The four-release doubling time is 8.0 and 8.3 months under these exclusions and 8.5 months with equal paper weights. The releases are the top tier of each lab; taking instead every release that raised the best horizon so far adds Sonnet 5 and gives 8.2 months ($[7.0, 9.8]$). In every one of these variants, dropping GPT-5 shortens the doubling time to about three months, so the trend depends mainly on the gap between the earliest release and the rest; the paper-bootstrap intervals do not include uncertainty from the choice of releases.

Field horizons and the range of the data. The largest problems per field have 38 (statistical mechanics), 130 (high-energy theory), 36 (complexity) and 63 (combinatorics) steps. In statistical mechanics the fitted crossing lies beyond 38 steps for Opus 5.5, GPT-6 Astra, GPT-5.6 Sol, Sonnet 5 and GPT-5.6 Luna (46 to 120 steps), and GPT-5 nano's crossings lie below one step in three fields; these values are extrapolations of the fitted curve and are marked in Figure 3c and Table 7. All other field horizons lie within the observed range, and the field ordering in the main text holds within it.

Calibration of the fitted curve. Against binned means, the fitted curve is within 0.08 of the data up to 31 steps for every capable model but underestimates the mean score above 32 steps by 0.13 to 0.20: the decline flattens at the largest sizes (Appendix K), so horizons that lie beyond the data are uncertain in both directions.

Table 7: **Research horizon by model and field** (fitted expected score $= 0.5$; steps; 95% paper-bootstrap intervals). Field fits use five papers each; * marks crossings outside the range of problem sizes observed in that field (largest problem: 38, 130, 36 and 63 steps), which are extrapolations.

Model	H_{50} [95% CI]	slope β [CI]	stat-mech	hep-th	cs.CC	math.CO
Opus 5.5	38.1 [23.0, 61.7]	-0.65 [-0.74, -0.58]	119.9* [89.7, 151.9]	91.2 [41.0, 155.1]	12.3 [10.8, 15.7]	20.6 [14.6, 25.9]
GPT-6 Astra	28.3 [20.3, 40.2]	-0.66 [-0.75, -0.59]	55.5* [27.8, 97.5]	52.8 [41.6, 134.7]	13.0 [9.3, 18.2]	17.2 [11.7, 23.4]
GPT-5.6 Sol	20.9 [15.4, 27.6]	-0.74 [-0.83, -0.67]	47.9* [30.8, 70.2]	30.4 [22.8, 44.2]	9.5 [8.5, 11.5]	14.2 [9.8, 18.4]
Sonnet 5	17.6 [11.9, 26.6]	-0.74 [-0.87, -0.64]	50.1* [30.8, 82.8]	32.1 [18.9, 46.6]	6.6 [5.2, 8.7]	10.7 [8.9, 13.6]
GPT-5.6 Terra	16.6 [12.3, 21.5]	-0.75 [-0.85, -0.69]	35.9 [31.5, 40.5]	21.9 [17.1, 31.9]	7.7 [6.6, 9.7]	12.4 [8.8, 16.1]
GPT-5.6 Luna	16.5 [12.7, 22.1]	-0.74 [-0.83, -0.66]	46.4* [33.5, 58.2]	19.7 [16.2, 39.7]	7.8 [6.8, 10.4]	11.9 [8.3, 15.6]
GPT-5 mini	9.8 [7.5, 12.9]	-0.69 [-0.79, -0.61]	22.8 [15.4, 32.6]	13.3 [10.7, 21.0]	4.4 [4.0, 5.1]	6.7 [5.9, 7.4]
GPT-5	10.0 [7.4, 13.3]	-0.66 [-0.76, -0.60]	23.0 [14.0, 35.8]	12.9 [8.3, 17.2]	4.5 [3.5, 6.0]	7.2 [5.7, 8.5]
Haiku 4.5	4.5 [3.5, 5.6]	-0.53 [-0.62, -0.47]	9.0 [5.4, 12.9]	5.4 [4.6, 6.0]	2.7 [2.4, 3.2]	3.2 [2.7, 3.6]
GPT-5 nano	0.7 [0.4, 1.0]	-0.39 [-0.50, -0.32]	0.5* [0.1, 1.1]	0.7* [0.2, 2.2]	1.0 [0.6, 1.2]	0.7* [0.5, 1.1]

H.2 THRESHOLDED VARIANT H_{50}^{τ}

For comparison with time horizons we also fit the probability of success, $\Pr[s(n) \geq \tau] = \sigma(\alpha' + \beta' \log_2 |\mathcal{L}(n)|)$, whose one-half crossing we write H_{50}^{τ}. At $\tau = 0.7$ it gives horizons about 0.5 to 0.6 times as long as H_{50} for the capable models and ranks the models almost the same way (Sonnet 5 and GPT-5.6 Terra swap). The per-paper, calibration, granularity and disputed-paper analyses in this subsection were computed for this variant; the granularity check on whole-paper scores does not depend on how the horizon is defined. These analyses use a pooled logistic fit of $\mathbf{1}[s(n) \geq \tau]$ on $\log_2$ node size over the 1,079 nodes of the 20 papers, with 2,000 paired paper-bootstrap draws (the same draws for every model, threshold and trend, so that ratios and doubling times are paired). Figure 11 shows how H_{50}^{τ} moves with the threshold and how well the logistic describes the data.

Thresholded variant: choice of threshold. H_{50}^{τ} falls with τ for every model except a 0.003-step rise for GPT-5 nano between $\tau = 0.6$ and 0.7 (Opus 5.5: 54, 44, 24, 11 and 4.1 steps at $\tau = 0.5, \ldots, 0.9$; GPT-5 nano: 0.99 to 0.74), and the ordering is nearly invariant: Kendall τ_b against the $\tau = 0.7$ ranking is 0.91–0.96 at the other thresholds, with at most two reversals, all among the closely spaced middle models; Opus 5.5 $>$ GPT-6 Astra $>$ GPT-5.6 Sol at every threshold. The threshold-free H_{50} of the main text gives a similar ordering (Kendall $\tau_b = 0.96$, $[0.82, 1.00]$; the only reversal is Sonnet 5 against GPT-5.6 Terra). Horizons normalised to each model's own midpoint between R_{full} and R_{root} measure a different construct: under them Astra overtakes Opus 5.5 and the leading-release trend stretches to 14–26 months per doubling, so we do not use model-specific thresholds.

Calibration. The logistic is a coarse summary of a curve that bends slightly. No model fails a Holm-corrected paper-bootstrap calibration test, but a quadratic in log size improves held-out (leave-one-paper-out) deviance for Opus 5.5 (-11.9, CI $[-21.9, -1.7]$) and not decisively for any other model; the largest gap is at 33–64 steps, where Opus 5.5 succeeds on 0.54 of nodes ($[0.21, 0.85]$, 13 nodes in 11 papers) against a fitted 0.30. Among single size measures, $\log_2$ of the number of steps has the lowest AIC for 8 of 10 models; adding paper fixed effects moves H_{50}^{τ} of Opus 5.5 from 23.7 to 21.6 steps.

Per-paper fits. Fitting each paper separately is not identified for the strongest models on 8 of the 20 papers each: Opus 5.5 succeeds on every node of 6 papers and its fit is separated on 2 more, GPT-6 Astra has 4 all-success and 4 separated papers, so those papers have no finite horizon; the median H_{50}^{τ} over the 12 papers with a finite fit is 12.2 steps ($[5.1, 86.6]$) for Opus 5.5. We therefore use the pooled fit throughout and give the per-paper values, with the papers whose fit is degenerate, in Figure 10.

Granularity. A step is the builder's unit, and papers were cut into 15 to 130 steps. We detect no association between whole-paper scores and the number of steps (Spearman between -0.29 and $+0.32$, all $p > 0.18$; $|\rho| \leq 0.06$ for Opus 5.5, GPT-6 Astra, GPT-5.6 Sol and Sonnet 5). Among finite fits, horizons correlate with tree size only for Opus 5.5 ($\rho = 0.64$, $p = 0.02$, 12 papers). This comparison is affected by non-identification: the other eight papers, mostly passed at every node, have smaller trees (median 27 against 37 steps). Paper fixed effects or equal paper weights change each horizon by at most 22% (Opus 5.5: 23.7, 21.6, 18.6 steps), preserving the top three and bottom four but changing the order of Terra, Sonnet 5 and Luna. Because a step is the builder's unit, horizons compare models on a common set of trees, not across differently built trees; their absolute scale is relative to this builder, and converting it to human working time would require timing experts on a sample of nodes.

Doubling time of the thresholded horizon. The four leading releases (GPT-5, GPT-5.6 Sol, GPT-6 Astra, Opus 5.5) give 6.1 months per doubling of $H_{50}^{0.7}$ ($[4.9, 7.9]$, $R^2 = 0.95$); the OpenAI series alone 6.8 ($[5.5, 8.5]$); the mid tier (GPT-5 mini, Sonnet 5, Terra) 9.7 ($[8.0, 12.2]$). Across thresholds 0.5–0.9 the estimate is 7.9, 5.8, 6.1, 8.4 and 12.9 months; leaving out any one paper gives 5.8–6.6 months; leaving out GPT-5 gives 2.6 months and leaving out any other release 5.9–6.8; weighting papers equally in the fit gives 6.7 months and excluding the two papers with earlier-builder trees 6.4. Two-point successions (GPT-5 to Sol 7.7 months, Sol to Astra 3.2 with CI $[1.8, 16.1]$) show that the trend is not uniform within the series. All intervals condition on the chosen releases, trees, judge and logistic form; they do not describe a population of models.

Disputed-paper sensitivity. For the thresholded horizon $H_{50}^{0.7}$, with the same four leading releases, we remove the complete trees of the audited papers (Appendix C); the same check for H_{50} is reported above. With 2,000 paired paper-bootstrap draws per cohort, excluding the two misstated roots' papers gives a doubling time of 6.40 months ($[4.99, 8.07]$); excluding all three disputed papers gives 6.42 ($[4.99, 8.23]$). On the latter subset, Opus 5.5's horizon is 2.09 times Sonnet 5's ($[1.57, 2.74]$) but only 1.11 times Astra's ($[0.84, 1.37]$). The selected-release trend and the separation from Sonnet 5 persist under these exclusions; an Opus–Astra horizon difference is not resolved. Removing complete trees also changes the size and field mixture, so this is a post-hoc diagnostic rather than a corrected horizon estimate or validation of the remaining problems.

I MEASUREMENT 2: DECOMPOSITION CURVE DETAILS

Figure 12 shows two properties of the decomposition curves that the main text states briefly. (a) For every cut size k we compute the best valid cut of k pieces by dynamic programming over the tree and compare it with the average cut of the same size along the deepest-first collapse: the best decomposition gains only 0.02–0.07 in $R(F)$, so the builder's decomposition is close to the most helpful one and the curves do not depend much on how the tree is collapsed. (b) Composing is costly and decomposing rarely hurts: 83–88% of the six strongest models' parents score below their weakest child, and only 1–3% score above the weighted mean of their children (GPT-5: 7%; GPT-5 nano, whose leaves often fail: 29%). The per-level shares behind Figure 4e and the per-paper deficits are released with the analysis outputs.

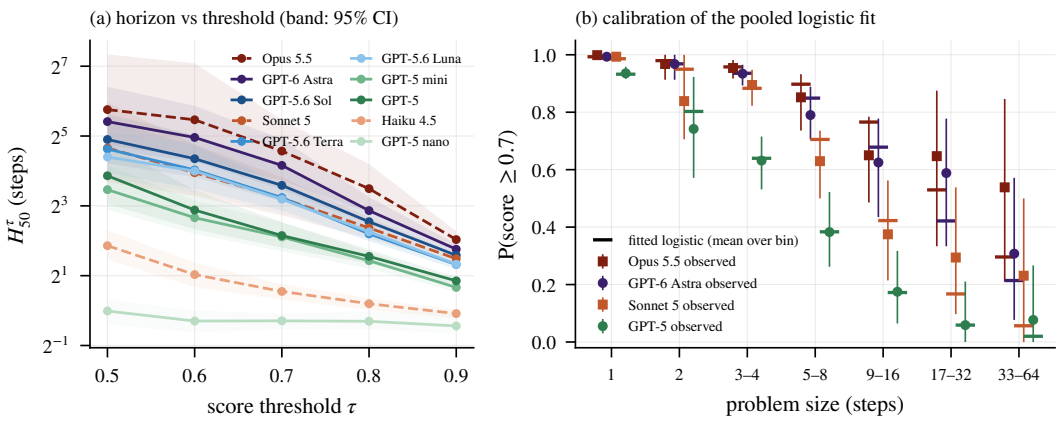

Figure 11: **Thresholded horizon.** (a) H_{50}^{τ} against the success threshold τ (pooled fit; bands are 95% paper-bootstrap intervals). (b) Calibration of the pooled logistic fit at $\tau = 0.7$ for four models: observed success rate per size bin with 95% paper-bootstrap interval (markers) against the mean fitted probability over the nodes in the bin (dashes).

The decomposition deficit. D is one minus the integral of the decomposition curve over the fraction decomposed $x \in [0, 1]$ (trapezoid rule on a 41-point grid) (deepest-first orders, averaged over ten orders), computed per paper and averaged over papers; it is the area above the curve in Figure 4a. It needs no success threshold and is finite for every model. We compared it with the number of pieces needed to reach a paper-level score τ (Table 8). The pieces needed are intuitive: at $\tau = 0.8$, Opus 5.5 needs four pieces on a typical paper, GPT-6 Astra and GPT-5.6 Sol about eight and a half, Sonnet 5 and the smaller GPT-5.6 models nine to eleven and GPT-5 fifteen and a half, while Haiku 4.5 reaches 0.8 on only 7 papers and GPT-5 nano on none. But the medians move in integer-like jumps, the release trend changes with τ (halving times from 5 to 20 months), and for some thresholds it moves by a factor of several when one release is dropped. Pieces read off a fitted logistic give tighter trends, but they depend on the chosen τ and on the logistic form, and for the weakest models on extrapolation far beyond the tree. D needs neither a threshold nor a functional form, is a direct average of the measured curve, is finite for every model and gives a stable trend, so it is the summary we use. By field, the six strongest models have $D = 0.09$ in statistical mechanics, 0.12 in high-energy theory, 0.18 in combinatorics and 0.20 in complexity.

Deficit over time. Over the four leading releases (GPT-5, GPT-5.6 Sol, GPT-6 Astra, Opus 5.5), a least-squares fit of $\log_2 D$ on release month gives a halving time of 12.3 months (95% paper-bootstrap interval $[10.7, 14.5]$, 2,000 draws, all with a negative slope; $R^2 = 0.98$). Dropping one release gives 9.0 (without GPT-5) to 13.1 months (without Opus 5.5), the same sensitivity to the earliest release as the horizon trend.

Table 8: **Choice of the Measurement-2 summary.** Trend of each candidate over the four frontier releases: R^2 of the log-linear fit, halving time in months with a 95% paper-bootstrap interval, range of halving times when one release is dropped, and the number of models for which the summary is finite. Pieces: median over papers of the smallest frontier with $R(F) \geq \tau$ (infinite if not reached); fitted pieces: from a logistic fit of $R(F)$ on $\log_2 |F|$ pooled over papers. D uses means over papers; the thresholded rows use medians and 1,000 draws, D 2,000.

summary	R^2	halving (months)	drop one release	models
pieces, $\tau = 0.6$	0.72	5.8 [4.1, 11.3]	1.4–8.1	9/10
pieces, $\tau = 0.7$	0.99	10.1 [4.3, 14.2]	6.8–10.4	9/10
pieces, $\tau = 0.8$	0.69	9.5 [6.8, 20.5]	2.8–14.1	8/10
pieces, $\tau = 0.9$	0.99	20.0 [9.1, 31.5]	14.4–21.1	8/10
fitted pieces, $\tau = 0.8$	0.99	6.0 [5.0, 7.9]	4.6–6.3	10/10
fitted pieces, $\tau = 0.9$	0.98	5.3 [4.4, 7.7]	5.0–5.5	10/10
deficit D (used)	0.98	12.3 [10.7, 14.5]	9.0–13.1	10/10

Half-gap size and merge order. An earlier summary of the decomposition curve is $N_{1/2}$, the largest cut whose score is at or below the midpoint of R_{full} and R_{root}. It has a median over papers of 4–6 nodes of the cut for the six strongest models (14–20% of a tree's leaves) and 7.5–9.5 for GPT-5, GPT-5 mini, Haiku 4.5 and GPT-5 nano (19–35%); Table 10 gives paper-bootstrap intervals of the medians. Collapsing sibling groups in a fully random order instead of deepest-first changes the median by at most one node for nine models (nano: 9.5 to 11), so the curves are robust to the merge order in the aggregate, although individual papers move by several nodes; we ran this merge-order check on $N_{1/2}$ and have not repeated it for D.

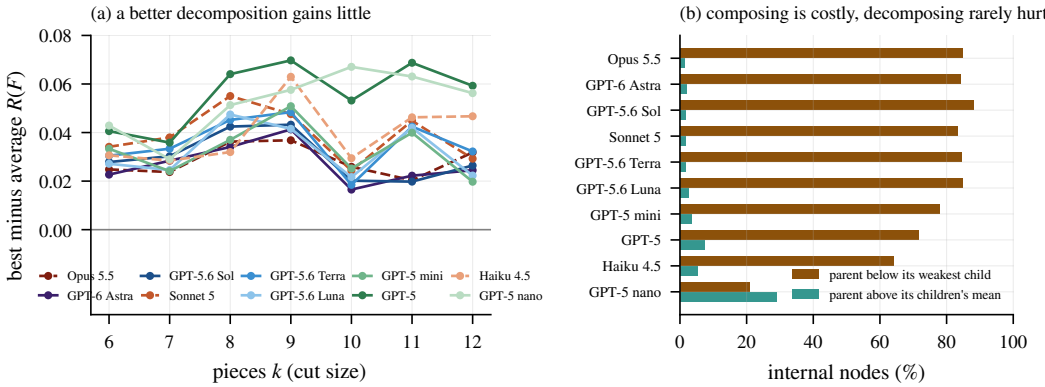

Figure 12: **Two further decomposition features.** (a) Best minus average $R(F)$ at each cut size, mean over papers (sizes reached on at least 10 papers). (b) Share of internal nodes scoring below their weakest child (brown) and above the weighted mean of their children (teal).

J FIELDS

Table 9 gives the whole-paper score of every model in every field, and Figure 13 the decomposition trajectories of every model by field. With five papers per field, we test for field effects on curve features averaged over the six strongest models (Opus 5.5, GPT-6 Astra, GPT-5.6 Sol, Sonnet 5, Terra, Luna) with permutation F-tests over the field labels (5,000 permutations, $p = (b+1)/(B+1)$). The field explains 63% of the between-paper variance in whole-paper score ($F = 9.0$, $p = 0.003$; an in-sample association without an interval) and 64% of the variance in the composition gap ($p = 0.001$), two strongly dependent tests; step scores are 0.96–0.97 in every field, so the spread there is one fortieth of the spread at the root (0.01 against 0.42). Grouping into physics (stat-mech, hep-th) and proof fields (cs.CC, math.CO) gives whole-paper scores of 0.65 against 0.34 (Mann–Whitney $p = 0.002$). The two fields within each group are not separated with five papers each: stat-mech vs. hep-th differ by 0.12 (Hedges $g = 0.7$), cs.CC vs. math.CO by 0.11 ($g = 0.8$); these are effect estimates, not evidence of equivalence. Level-free shape features (the area under the decomposition curve after rescaling it to $[R_{\text{root}}, R_{\text{full}}]$, and the fraction of the tree that must be decomposed to close half the gap) are not resolved by field with this sample ($p = 0.07$ and $p = 0.18$); the one nominal pairwise difference, a larger rescaled area for math.CO than for cs.CC papers ($g = -1.6$, $p = 0.03$ on the area, uncorrected), is suggestive with $n = 5$. As an exploratory check, a nearest-centroid classifier with one paper held out recovers the field of a paper from its full decomposition curve 70% of the time (chance 25%, permutation $p < 0.001$, whole-paper labels permuted and the classifier refitted each time) and physics vs. proof 80% of the time ($p = 0.01$), but from its rescaled shape alone only 35% ($p = 0.22$).

K WHERE THE LOST SCORE GOES: DETAILS

Figure 14 shows the mean node score at each depth of the tree. This section gives the analyses behind Figure 5; Table 10 summarises them by model. Intervals resample papers (2,000 draws); the size regressions include paper fixed effects.

Table 9: **Whole-paper scores by field.** Each cell averages the importance-weighted root score over the same five papers.

Model	stat-mech	hep-th	cs.CC	math.CO
Opus 5.5	0.83	0.75	0.51	0.40
GPT-6 Astra	0.64	0.60	0.47	0.34
GPT-5.6 Sol	0.69	0.60	0.39	0.29
Sonnet 5	0.73	0.60	0.31	0.24
GPT-5.6 Terra	0.69	0.48	0.36	0.24
GPT-5.6 Luna	0.67	0.50	0.35	0.25
GPT-5 mini	0.56	0.44	0.22	0.19
GPT-5	0.49	0.47	0.25	0.19
Haiku 4.5	0.39	0.31	0.17	0.16
GPT-5 nano	0.17	0.15	0.07	0.10

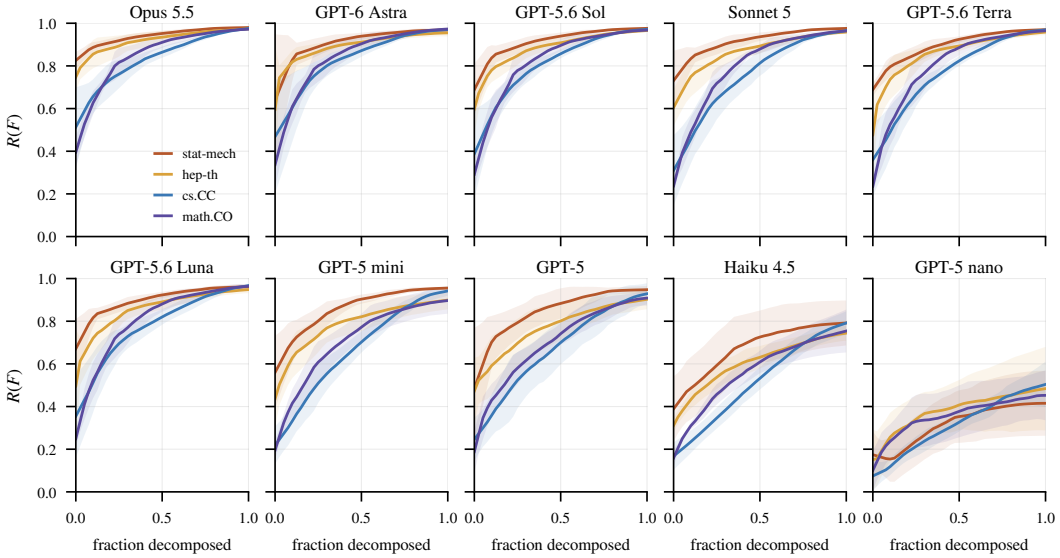

Figure 13: **Decomposition curves by field and model.** Decomposition curves, with the whole-paper problem at $x = 0$ and isolated atomic steps at $x = 1$; each field contains five papers.

Score and size. For every model the logit of the clipped continuous node score has a negative linear association with $\log_2$ size (slopes -0.45 to -0.87 per doubling, every interval below zero), and a positive quadratic term captures flattening at large sizes; the quadratic improves leave-one-paper-out error for 8 of 10 models. The slope is not universal: Sonnet 5 loses 0.14 logits per doubling more than Opus 5.5 ($[0.12, 0.18]$). With 19 nodes of 16–32 leaves and 14 above 32, the flattening is measured but its asymptote is not.

Composition tax. The tax of a parent, $\text{logit}\, s(\text{parent}) - \text{mean}_{\text{children}} \text{logit}\, s(\text{child})$, is -1.3 to -1.7 logits for the eight models above 0.9 at the leaves and shows no clear depth trend (Opus 5.5: -0.97 at the root, -1.2 to -1.3 at depths 1–5). For Opus 5.5, 85% of parents ($[81, 88]$) score below their weakest child; the mean and the minimum of the children's scores are associated with the parent about equally in-sample (R^2 0.52 vs. 0.48), so the data do not single out a weakest-link mechanism. GPT-5 nano is the exception (21% of parents below the weakest child): when the leaves themselves fail, there is little left to lose in composition.

Nesting. With success defined as $s(n) \geq 0.7$, the models are close to nested: Opus 5.5 succeeds on 99.6% of the nodes Sonnet 5 succeeds on ($[99.0, 100]$), and nodes that only Sonnet 5 solves are 0.4% of all nodes ($[0.0, 0.9]$); across all 45 model pairs the weaker-only share is at most 6.5% (GPT-5 mini vs. GPT-5, whose whole-paper scores differ by 0.004). Node-level Spearman correlations between models are 0.21–0.82 over all nodes but only 0.08–0.48 over leaves, where ceiling effects

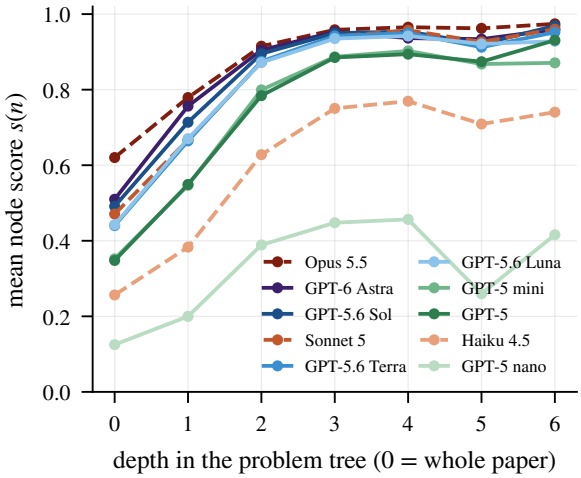

Figure 14: **Score by depth.** Mean weighted node score $s(n)$ at each depth (0 = the whole paper), averaged within and then over papers; depth-0 points are the whole-paper scores of Table 1. The six strongest models score 0.87–0.92 at depth two and 0.94–0.96 at depth three; depths 4, 5 and 6 occur in 17, 7 and 4 papers only, so the profile mixes papers as well as granularities.

dominate; the shared size gradient inflates the pooled correlations, so this is nesting of success, not evidence of a one-dimensional difficulty scale.

Where newer models improve. Conditioning on the nodes an earlier release fails, the fraction its successor fixes falls with node size: GPT-5 to GPT-5.6 Sol fixes 100% of failed leaves and 15% of failed nodes above 32 leaves ($[0, 43]$); Haiku 4.5 to Sonnet 5, 98% and 21%; Sonnet 5 to Opus 5.5, 100% and 36% ($[0, 67]$); GPT-5 nano to GPT-5.6 Luna, the third pair in Figure 5c, 99% and 14% ($[0, 39]$). The leaf-minus-large-node difference is 54–88 percentage points for the other comparisons; Sol to Astra is an exception, with zero of one failed leaf and one of eleven failed large nodes fixed, too little leaf support for a stable comparison.

Composition versus steps. For each model the whole-paper deficit against Opus 5.5 splits exactly, paper by paper, into an atomic-step deficit (R_{full} difference) and an extra composition gap. Composition accounts for 92–96% of the deficit of GPT-6 Astra, GPT-5.6 Sol, Sonnet 5 and GPT-5.6 Terra and 92% of GPT-5.6 Luna's (e.g. Sonnet 5: 15.0 points, $[10.6, 20.0]$, of which 0.9 atomic, $[0.6, 1.3]$, and 14.0 composition, $[9.6, 19.0]$; Astra: 0.9 $[0.3, 1.5]$ and 10.2 $[1.6, 20.6]$; Sol: 0.5 $[0.3, 0.7]$ and 12.4 $[8.6, 16.7]$), 80% of GPT-5's and GPT-5 mini's, 44% of Haiku 4.5's, and none of GPT-5 nano's: its 49.5-point deficit is atomic (51 points; extra composition gap -1.6, $[-11.8, +9.4]$). Reading the same thing off depth instead of the endpoints gives the same picture: Sonnet 5 trails Opus 5.5 by 13.2 points at depths 0–1 and by 2.6 points at depths ≥ 3.

Answer length. A whole-paper answer is graded against far more reference text than a single step (median root answer 0.1 to 0.3 times the reference text beneath the root for the capable models), so a fixed-length answer spread over more references could produce a size law on its own: if a model's odds of covering a reference were inversely proportional to the number of references, the logit slope per doubling would be $-\ln 2 \approx -0.69$, close to the fitted slopes. The saved answers do not support this reading. Models write longer answers for larger problems (length grows as size to the power 0.14 to 0.34 for the eight capable models; Opus 5.5 from a median of 9.6k characters on single steps to 34k above 32 steps). Regressing the logit of the node score on $\log_2$ size and $\log_2$ answer length with paper fixed effects, a fixed-length dilution would give a length coefficient near $+0.69$; for the strong models the coefficient is instead small and mostly negative (-0.31 to -0.09, Sonnet 5 $+0.27$ with an interval including zero), and adding length changes the size coefficient by at most 0.19 (Opus 5.5: -0.67 to -0.61). GPT-5 mini's coefficient is -0.70; only GPT-5 and GPT-5 nano show a large positive one ($+0.66$, $+0.71$), because their failed answers are short. This rules out answer length as the explanation within the answers we have, but it does not test whether a model given an

explicit, larger output budget for a whole paper would close the gap; that requires a matched-budget run we have not done.

Leaf difficulty is local. A sequential sum-of-squares decomposition of leaf scores (field, then paper, then top-level branch, then sibling group, then residual) attributes 19% ($[13, 25]$) of within-model variation to field, paper and branch together and 81% to sibling groups and individual leaves: in this in-sample partition, which step a model misses is mostly local to sibling groups and leaves.

Table 10: **Anatomy of the composition gap by model.** Size-law slope (fixed-effects logit regression on $\log_2$ size), mean composition tax, share of parents scoring below their weakest child, median $N_{1/2}$ over papers on the decomposition curves (deepest-first orders; unit: frontier problems) and its median fraction of the leaves L; 95% paper-bootstrap intervals.

Model	Slope (logit/doubling)	Mean tax (logit)	Parents below weakest child (%)	Median $N_{1/2}$ (nodes of the cut)	$N_{1/2}/L$
Opus 5.5	-0.67 [-0.76, -0.59]	-1.26 [-1.40, -1.14]	85 [81, 88]	5.0 [3.0, 7.5]	0.16
GPT-6 Astra	-0.70 [-0.79, -0.64]	-1.29 [-1.42, -1.19]	84 [81, 88]	4.0 [3.0, 7.5]	0.16
GPT-5.6 Sol	-0.78 [-0.86, -0.71]	-1.45 [-1.54, -1.36]	88 [86, 91]	4.5 [3.5, 6.0]	0.14
Sonnet 5	-0.81 [-0.93, -0.72]	-1.53 [-1.72, -1.37]	83 [80, 86]	6.0 [5.0, 8.0]	0.20
GPT-5.6 Terra	-0.83 [-0.91, -0.77]	-1.55 [-1.66, -1.47]	84 [80, 89]	5.0 [4.0, 6.5]	0.18
GPT-5.6 Luna	-0.82 [-0.90, -0.75]	-1.52 [-1.63, -1.42]	85 [82, 88]	5.0 [3.0, 7.0]	0.18
GPT-5 mini	-0.84 [-0.94, -0.77]	-1.62 [-1.78, -1.48]	78 [73, 83]	8.0 [5.5, 9.0]	0.19
GPT-5	-0.87 [-0.98, -0.79]	-1.74 [-1.89, -1.60]	72 [67, 76]	7.5 [7.0, 9.0]	0.22
Haiku 4.5	-0.68 [-0.78, -0.61]	-1.37 [-1.53, -1.24]	64 [59, 71]	9.0 [7.5, 11.0]	0.31
GPT-5 nano	-0.45 [-0.61, -0.34]	-0.90 [-1.21, -0.64]	21 [15, 27]	9.5 [6.5, 12.5]	0.35

What the judge rationales report. From the 16,513 eligible low-scoring (< 0.8) judgments of a proper ancestor against one of its leaves (recorded model failures and answers under 500 characters excluded), we drew 600 by stratified random sampling over model, depth of the judged node and field, and 100 high-score controls; one LLM reader labelled each judge rationale with a primary cause from a codebook fixed before sampling, and inspected the answer and leaf solution for 154 pairs, including every proposed alternative route and judge error. With sampling weights and paper-cluster bootstrap intervals (5,000 draws, stratified by field), the primary causes of the low scores are an omitted result (45%, $[40, 53]$), an omitted derivation of a stated result (18%), a scope mismatch (13%), a wrong result or invalid step (15%, $[11, 20]$), a contradiction of the leaf (1%), a supported alternative route (6%) and a confirmed judge misreading (1.5%). Cases with no identified asserted error make up 82.9% ($[77.1, 87.0]$) of the low-scoring judgments and 81.9% ($[75.6, 86.4]$) of the lost score; explicit errors, in either label position, 16.6% ($[11.9, 23.1]$) of the lost score, rising to 20.8% if the ten unresolved cases are all counted as errors. The error share of lost score is 13.9% at the root and 23.5% at depths of three or more, a difference of $+9.6$ points that is not separated from zero ($[-0.6, +20.2]$) and is confounded with node size. Supported alternative-route penalties occur in 6.5% ($[2.5, 11.0]$) of low-scoring root judgments (2.3% with high-confidence route labels only); this is the incidence of the route-coverage penalty, not the size of a score correction. The audits confirmed 15 judge misreadings; because inspection targeted flagged cases, this is not an estimate of the judge's error rate. All labels come from a single LLM reader without human adjudication, describe the judge's diagnosis rather than independent correctness, and do not show that an omission reflects an inability to compose; the intervals cover sampling of papers, not labelling uncertainty.

L MULTIPLE COMPARISONS

Table 1 reports an interval for each comparison separately, so it does not certify a complete ranking. As an exploratory multiplicity check, we jointly resample papers across all models in 20,000 new bootstrap draws (seed 20260925); small differences from the original pointwise intervals reflect Monte Carlo variation. For the nine differences d_j against Sonnet 5, let $\hat{d}_j$ be the observed mean and $\hat{\sigma}_j$ the bootstrap standard deviation of that mean. The 95th percentile c of $\max_j |(d_j^* - \hat{d}_j)/\hat{\sigma}_j|$ gives the approximate simultaneous intervals $\hat{d}_j \pm c\hat{\sigma}_j$. This controls the stated contrast family approximately under the bootstrap, not every pairwise comparison or analysis in the paper. The scales are fixed across resamples; these are not resample-studentized intervals.

Table 11: **Multiplicity check on whole-paper scores.** Paired differences against Sonnet 5, with simultaneous 95% intervals over the nine displayed contrasts.

Model	Difference	Simultaneous interval
Opus 5.5	$+0.150$	$[+0.086, +0.213]$
GPT-6 Astra	$+0.039$	$[-0.087, +0.165]$
GPT-5.6 Sol	$+0.021$	$[-0.024, +0.066]$
GPT-5.6 Terra	-0.031	$[-0.115, +0.054]$
GPT-5.6 Luna	-0.028	$[-0.107, +0.050]$
GPT-5 mini	-0.119	$[-0.170, -0.067]$
GPT-5	-0.123	$[-0.232, -0.014]$
Haiku 4.5	-0.214	$[-0.294, -0.134]$
GPT-5 nano	-0.346	$[-0.491, -0.201]$

M MODEL BEHAVIOUR AND FAILURES

All models received the same instructions and the same Python tool; Table 12 shows how they used it. Claude models ran Python one to three times per problem on average, while the newer OpenAI models rarely used it and reasoned internally instead. The same "high" reasoning setting meant about 0.9–4.3k reasoning tokens per problem for the GPT-5.6 and GPT-6 models but 10–13k for GPT-5 and GPT-5 mini.

Answer length. Longer answers are not simply better: after removing paper effects, depth and node size, the partial correlation between answer length and score is 0.79 for GPT-5 nano over all its answers but 0.03 ($[-0.09, +0.14]$) once its short and failed answers are excluded, and 0.03 ($[-0.14, +0.20]$) for Opus 5.5 on internal nodes; only GPT-6 Astra (0.30, $[0.09, 0.46]$) and GPT-5.6 Luna (0.35) show a positive association on internal nodes. Reasoning-token use on the 140 OpenAI root solves has a partial correlation of 0.11 ($[-0.08, +0.29]$) with the root score. These are observational associations with ceiling effects and endogenous effort; no randomised effort comparison was run.

Failures and retries. Answers that were short, corrupted by the harness's shell commands, or lost to an infrastructure failure received one retry, and the affected lineages were re-judged; answers that hit the output-length cap were not retried. The two harnesses detected short answers slightly differently (OpenAI: under 2,000 characters; Claude: at most 2,000 bytes without an explicit no-answer marker). In total 718 OpenAI-model nodes were retried once (five of them twice, by accident); one Opus 5.5 node, an interrupted file edit, was retried successfully (no final Opus 5.5 failures); Sonnet 5 kept three cap failures. GPT-5 kept 37 nodes (3.4%) with answers under 2,000 characters consisting mostly of shell-test text (some may be legitimate short answers); on the remaining nodes it scores 0.02 higher than GPT-5 mini. GPT-5 nano has no answer or a flagged short or corrupted answer on 46% of nodes (the two categories combined); on one node its shell commands overwrote the problem file, so its retry answered an empty prompt (recorded as a failure; every other problem file is byte-identical across models). We keep the judged scores rather than forcing failures to zero, since short or partial answers can earn some credit. Ten explicit no-answer records (eight from nano, two from Sonnet 5) nevertheless received positive scores, and we could not determine why; setting them to zero changes nano's whole-paper score by 0.00014 and no other score, and no affected node crosses 0.7.

N MODEL IDENTITY

Because hosted models can be routed to a different variant than requested, we verified the served model. For OpenAI models called through the API, every request passed through a logging proxy that records the model reported in the response; all requests were served as the requested model (or a dated snapshot of it). For OpenAI models accessed through a subscription, the model and reasoning effort of each response were read from the client's network trace; 80–90% of GPT-5.6 nodes and 87% of GPT-6 Astra nodes carry such a record (the rest used a transport that was not traced), all showing the requested model at high effort. For Claude models, every agent transcript records the

Table 12: **Model behaviour.** Python runs are means per node (Claude: per agent launch); reasoning tokens and answer lengths are medians per node; thinking tokens are reported by the OpenAI API only.

Model	Python calls mean/node	Reasoning tokens median/node	Answer chars median	Minutes/node median	Failures
Opus 5.5	1.0	–	10,862	–	0
Sonnet 5	2.6	–	18,772	–	3
Haiku 4.5	1.1	–	9,520	–	0
GPT-6 Astra	0.10	872	11,450	3.4	0
GPT-5.6 Sol	0.12	2,114	8,129	2.0	0
GPT-5.6 Terra	0.17	1,938	6,978	1.8	0
GPT-5.6 Luna	0.26	4,293	7,407	1.9	0
GPT-5	0.31	9,984	8,581	5.8	2
GPT-5 mini	0.94	13,120	12,359	3.6	0
GPT-5 nano	0.08	9,216	3,899	9.3	159

model of every message; all 1,137 Opus 5.5 model messages came from `claude-opus-5-5`, and each job probed the served model before solving.

O COST

List-price cost per node (model only): Opus 5.5 $0.77, Sonnet 5 $1.27, Haiku 4.5 $0.28, GPT-6 Astra $0.51, GPT-5.6 Sol $0.29, Terra $0.15, Luna $0.09, GPT-5 $0.25, mini $0.05, nano $0.01. Recorded judge usage totals about $5.3k and API usage about $0.9k at list prices, alongside subscription usage; because some attempt records were overwritten, these are lower bounds rather than a complete bill. Building a tree cost $11–$54 at list prices for the 18 trees from the current builder.