Position: The Window for Preventive Action Before AI Saturates Most Cognitive Benchmarks is Closing

Jérémy Andréoletti, Antoine Maier, Laura Domenech, Tom David

Position: The Window for Preventive Action Before AI Saturates Most Cognitive Benchmarks is Closing

A Preprint

Jérémy Andréoletti* **Antoine Maier** **Laura Domenech** **Tom David**

General-Purpose AI Policy Lab

Abstract

In this position paper, we argue that without coordinated intervention, AI systems will by default saturate most existing cognitive benchmarks by 2030, approaching or exceeding human expert performance across a broad range of measurable tasks. Using a joint hierarchical Bayesian model, we forecast frontier performance trajectories across 63 benchmarks spanning reasoning, mathematics, coding, scientific knowledge, and agentic capabilities. We find that nearly all current benchmarks (98% in our set, 80% CI: 97–100%) are on track to saturate within four years. This finding is robust to modeling choices: sensitivity analyses show that altering key assumptions yields qualitatively similar conclusions. Security-critical benchmarks (cybersecurity, autonomous AI R&D, biology, chemistry) show even faster trajectories, saturating before 2028. While benchmark saturation does not equate to artificial general intelligence, these findings add to a converging body of evidence indicating that superhuman performance on cognitive tasks is approaching faster than commonly assumed. Acknowledging that we neither understand nor control the behavior of current GPAI models, this timeline leaves limited runway before irreversible impacts. We outline implications for global governance and international coordination, AI safety research, and evaluation practices.

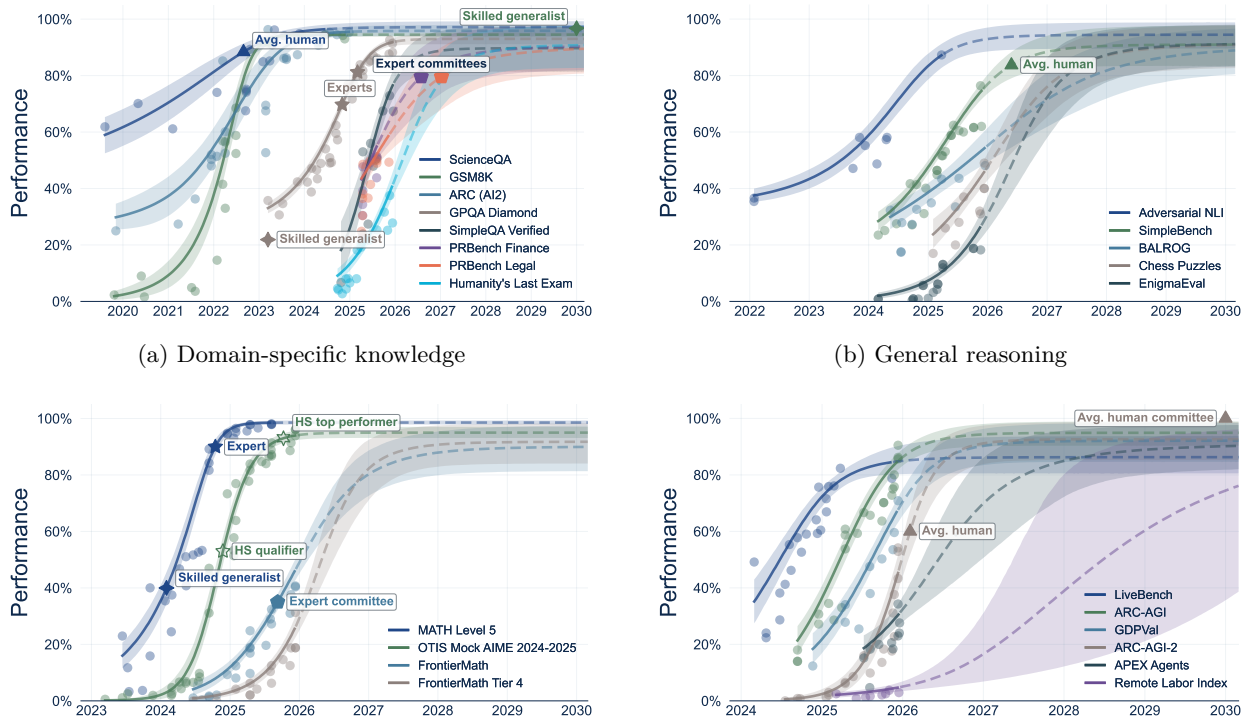
Keywords AI progress · benchmark saturation · capability forecasting · AI safety · AGI

1 Introduction

The pace at which new benchmarks have been developed over the past five years has only been matched by the speed at which these benchmarks have been climbed by each new generation of General-Purpose AI (GPAI) models (Epoch AI [2024a], Bengio et al. [2025], AISI [2025]; Figure 1). Tasks that were once considered out of reach, from doctorate-level science questions to competitive mathematics and autonomous software engineering, are now routinely solved by frontier models. Yet discussions about when AI might reach or exceed human-level performance on broad cognitive tasks often place it decades away [Grace et al., 2018, 2025] or treat it as deeply uncertain.

Our position: Based on a systematic analysis of benchmark trajectories, we argue that AI systems are on track to saturate most existing cognitive benchmarks by 2030, approaching or exceeding human expert performance across a broad range of measurable cognitive tasks. This projection, which includes security-critical capabilities, leaves a limited runway before irreversible impacts, thereby warranting urgent preventive action. This is not a prediction that AGI will necessarily arrive by 2030, as the relationship between benchmark performance and general intelligence remains contested [Chollet, 2019]. However, it is a call to take seriously the empirical trend that AI is saturating our best evaluations faster than anticipated, and faster than current coordination efforts for global GPAI risk management.

If correct, this timeline leaves little time to develop robust alignment or control techniques, given the cur-



(a) Domain-specific knowledge

(b) General reasoning

(c) Advanced mathematics

(d) AGI progress benchmarks

Figure 1: **Benchmark trajectories across capability categories.** Points show frontier model scores at release; solid lines show posterior median trajectories; shaded regions indicate 80% credible intervals. Dashed lines extrapolate to 2030. Human reference points are depicted by star-like symbols (representing average individual performance) and polygons (representing committee majority-vote aggregates). The number of vertices denotes the expertise level: 3 for average human, 4 for skilled generalist, 5 for domain expert, and 6 for top performer. A hollow symbol specifically denotes trained high-school level participants. See more details in Appendix E. All categories show rapid progress towards saturation, with several benchmarks already exceeding human expert baselines.

rent lack of any reliable method for understanding and steering powerful AI systems [Casper et al., 2023, Bengio et al., 2025, Maier et al., 2025, Ngo et al., 2025]. It also implies an urgency regarding international coordination on AI; the earlier multilateral discussions gain momentum, the more time will be available for countries to converge on a global course of action before GPAI models could start posing irreversible security issues. Even in optimistic safety scenarios, this pace of progress only leaves few years for institutions to adapt to transformative AI capabilities.

Our contribution is threefold: (1) We provide quantitative evidence from 63 benchmarks and 37 human baselines showing convergent saturation trajectories based on a new modeling framework; (2) We analyze AI progress specifically in security-critical capability domains; (3) We outline implications for researchers and policymakers.

2 Evidence: Benchmark Trajectories

2.1 Data and Methodology

We analyze performance trajectories on 63 benchmarks spanning diverse cognitive capabilities: commonsense, reasoning, scientific knowledge, mathematical problem-solving, code generation, agentic computer use, and multimodal understanding. Benchmark scores are sourced from the Epoch AI Benchmark database [Epoch AI, 2024a], Scale AI leaderboards [Scale AI, 2025], and evaluations by the RAND Corporation [Dev et al., 2025]. We exclude unsaturated benchmarks lacking data points from frontier models released after January 2025. To ensure that our projections are based on a well-defined observation window and can be reproduced, we restrict the fitting data to frontier models released before January 1, 2026.

To project future performance, we employ hierarchical Bayesian models which capture the S-shaped trajectories empirically observed in benchmark progress.

Our approach introduces three methodological improvements over existing AI capability forecasting:

(1) Asymmetric growth curves. We use Harvey curves [Harvey, 1984] instead of standard logistic functions. Unlike the logistic curve, the Harvey function allows for an asymmetry between the initial acceleration of benchmark progress and its deceleration as scores reach saturation (Figure 2). This asymmetry is captured by a shape parameter $\alpha > 1$, which reduces the Harvey curve to a logistic function when $\alpha = 2$. This choice avoids the logistic assumption that performance should accelerate and then decelerate at the same rate. A sensitivity analysis in Appendix B confirms that the main saturation finding is robust to removing this assumption.

(2) Hierarchical Bayesian modeling. Rather than fitting each benchmark independently, we jointly model all 63 benchmarks with shared hyperpriors over growth rates, upper asymptotes, as well as shape and noise parameters. This allows the exchange of information across benchmarks: data-rich benchmarks inform projections for newer or more data-sparse benchmarks. The cross-validation (LOO-ELPD) model comparison confirms that the joint model achieves better fit, correcting for the additional complexity, than independent model variants (Supp. Table 1), and the main saturation finding is robust to removing the hierarchical structure (Appendix B).

(3) Skewed likelihood at the frontier. Benchmark scores typically underestimate true model capabilities due to suboptimal prompting, scaffolding, or evaluation conditions (the "elicitation gap"). They also cannot estimate the capabilities of unreleased models. To mitigate this gap, we model observations at the top-3 frontier with a skew-normal likelihood, allowing scores to fall predominantly below a latent curve that represents the performance frontier. We verify that this modeling choice does not inflate our projections: replacing the skew-normal with a symmetric normal likelihood yields qualitatively similar saturation conclusions (Appendix B).

We validate our forecasting methodology through temporal holdout: training on data before January 2025 and evaluating predictions on observed scores in 2025. The skew-normal likelihood as well as the joint model consistently improve both CRPS and RMSE scores, while Harvey and logistic variants achieve similar predictive accuracy (Supp. Table 3). The more general Harvey curves are preferred in this article as they better align with our theoretical understanding of capability progress, and achieve the best cross-validation fit. Unlike classical machine learning, Bayesian inference regularizes through prior specifications, avoiding overfitting despite the model's greater complexity. All uncertainty intervals reported in this paper are Bayesian credible intervals (posterior quantiles). We additionally validate these intervals using conformal prediction, a distribution-free method, confirming that the Bayesian intervals achieve their nominal coverage (Supp. Table 3). See Appendix A for model specification, Appendix B for the sensitivity analysis and Appendix C for validation details.

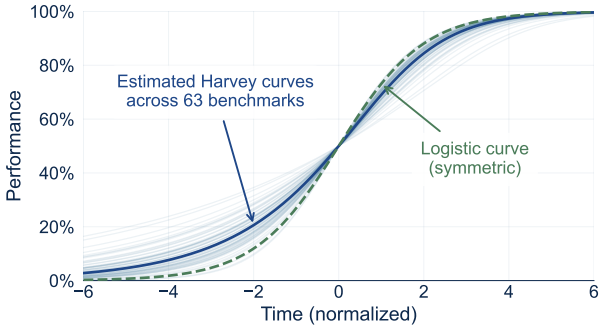


Figure 2: **Harvey curves capture asymmetric progress trajectories.** Fitted Harvey curves (blue) for 63 benchmarks show gradual acceleration followed by rapid saturation, compared to the symmetric logistic curve (green dashed). All curves are standardized to reach 50% performance at time 0 and have a growth rate of 1, in order to compare shapes only. The thick blue curve is the median Harvey trajectory.

2.2 Main Finding: Saturation by 2030

Our central empirical finding is that **approximately 98% of analyzed benchmarks are projected to reach saturation before 2030** (Figure 3), where saturation is defined as achieving 95% of their score range (between random chance and their estimated asymptote). This threshold serves as a proxy for a benchmark being essentially solved. In practice, benchmarks are typically considered saturated well before this point, and the smooth sigmoidal fit spreads the final percentage points over an extended period that does not reflect meaningful capability differences. Sensitivity analyses with thresholds of 90% and 99% confirm that this choice does not qualitatively affect our conclusions (Appendix B).

This finding is robust across benchmark categories, as shown in Figure 1. To contextualize these projections, we overlay human performance baselines where available. These baselines range from average crowdworkers to world-class experts, providing interpretable reference points for AI progress. Detailed human baseline sources are provided in Appendix E.

- **Domain-specific knowledge** (ScienceQA, ARC AI2, GSM8K, GPQA Diamond, PRBench, SimpleQA Verified, Humanity's Last Exam) [Lu et al., 2022, Clark et al., 2018, Cobbe et al., 2021, Rein et al., 2023, Wang et al., 2024, Akyürek et al., 2025, Haas

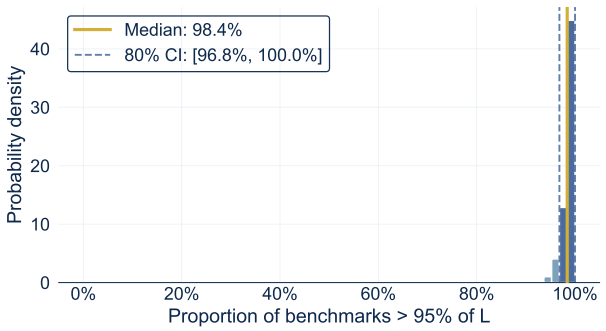


Figure 3: **Nearly all benchmarks projected to saturate by 2030.** Posterior distribution of the proportion of benchmarks reaching 95% of their maximum score range by 2030. The median is 98.4%, indicating that saturation of most benchmarks by 2030 is the default scenario, barring exogenous intervention.

et al., 2025, Phan et al., 2025]. Student to expert-level science questions in physics, chemistry, and biology; professional reasoning in law and finance; factual knowledge across disciplines. On GPQA Diamond, PhD-level experts achieve 69–81% accuracy [Rein et al., 2023]; frontier models now exceed this range. Saturation projected by 2029.

- **General reasoning** (Adversarial NLI, SimpleBench, BALROG, Chess Puzzles, EnigmaEval) [Nie et al., 2020, Philip and Hemang, 2024, Paglieri et al., 2025, Epoch AI, 2025a, Wang et al., 2025]. Spatial and temporal reasoning, multi-step logical deduction, strategic planning in games, and long multimodal reasoning challenges. Saturation projected by 2029 to 2030.
- **Mathematical reasoning** (MATH, OTIS Mock AIME, FrontierMath) [Hendrycks et al., 2021, Epoch AI, 2025b,c]. MATH and AIME contain problems from math competitions; FrontierMath Tier 4 consists of handcrafted questions aimed at taking hours for domain-expert mathematicians to solve. On MATH Level 5, a skilled human (PhD student) achieves 40%, while a domain expert (three-time IMO gold medalist) achieves 90% [Hendrycks et al., 2021]; frontier models now exceed the latter. On FrontierMath, teams of mathematicians collectively solve only 35% of problems in four and a half hours, with internet access [Epoch AI, 2025c]. Our analysis projects rapid progress and saturation in 2028.
- **AGI progress** (LiveBench, ARC-AGI v1 and v2, soon v3, Remote Labor Index, APEX-Agents, GDPval) [White et al., 2025, Chollet, 2019, ARC Prize, 2019, 2025a,b, Mazeika et al., 2025, Vidgen et al., 2026, Patwardhan et al., 2025]. Benchmarks designed to resist memorization and measure general intelligence. LiveBench uses regularly-updated questions spanning many domains. ARC-AGI tests the ability to learn new abstract concepts from few examples, which is currently easy for humans but difficult for AIs. Average humans achieve 77% on ARC-AGI-1, while STEM graduates and human panels reach 98% [ARC Prize, 2019]. The Remote Labor Index (RLI) measures completion rates on long real-world tasks from online freelance platforms. Similarly, APEX-Agents grades agents on ∼2-hour banking, consulting and law tasks in realistic file environments, while GDPval has experts compare AI deliverables to professional work on ∼7-hour research and writing tasks across 44 U.S. occupations. LiveBench, ARC-AGI-2 and GDPval should saturate soon, APEX-Agents by 2029; predictions for RLI are much more uncertain.

Benchmarks measuring capabilities with direct security implications show particularly rapid trajectories (Figure 4).

- **Cybersecurity** (TerminalBench, OSWorld, Cybench, TheAgentCompany, MCP Atlas) [Merrill et al., 2026, Xie et al., 2024, Zhang et al., 2025, Xu et al., 2025, Bandi et al., 2025]. Benchmarks evaluating agentic computer use, vulnerability discovery, and exploitation. Frontier models are progressing by tens of percentage points annually, with saturation projected by 2028.
- **AI R&D automation.** (Aider Polyglot, METR Time Horizons, WeirdML, SWE-Bench Verified, Bash-Only and Pro, GSO-Bench, PostTrainBench) [aider, 2024, Kwa et al., 2025, Ihle, 2025, Epoch AI, 2024b, Jimenez et al., 2024, Deng et al., 2025, Shetty et al., 2025, Rank et al., 2026]. Benchmarks measuring autonomous software engineering and ML research capabilities. Saturation projected by 2028 (except for PostTrainBench), suggesting that AI systems capable of significantly accelerating AI research may emerge in the coming years.
- **Expert biological and chemical knowledge.** (MMLU Pro, Weapons of Mass Destruction Proxy WMDP, GPQA Diamond, LAB-Bench, BioLP-bench) [Wang et al., 2024, Li et al., 2024a, Rein et al., 2023, Laurent et al., 2024, Dev et al., 2025] Benchmarks assessing scientific knowledge relevant

to biosecurity threats, including pathogen characteristics, synthesis procedures, and safety protocols, or evaluating practical laboratory skills such as protocol understanding, literature search, and experimental design. PhD-level domain experts achieve 38–83% on these benchmarks, depending on the specific task (Appendix E); frontier models already match or exceed these baselines on several sub-tasks. Saturation projected by 2027-2028.

Three additional benchmark categories are presented in Appendix D: **Commonsense reasoning** (OpenBookQA, HellaSwag, PIQA, WinoGrande) [Mihaylov et al., 2018, Zellers et al., 2019, Bisk et al., 2020, Sakaguchi et al., 2020]; **Language understanding and writing** (Lech Mazur Writing, Fiction.LiveBench, TutorBench, MultiChallenge, MultiNRC) [Mazur, 2026, Epoch AI, 2025d, Srinivasa et al., 2025, Sirdeshmukh et al., 2025, Fabbri et al., 2025] and **Multimodal understanding** (GeoBench, CAD-Eval, Visual Task Assessement VISTA, VPCT, Audio MultiChallenge, VisualToolBench) [ccmdi, 2026, Epoch AI, 2025e, Scale AI, 2025, Brower, 2025, Gosai et al., 2025, Guo et al., 2025].

Overall, all the projections we derived from the joint hierarchical Harvey model estimate benchmark saturation by 2028 to 2030. Uncertainty intervals widen for longer forecast horizons, but even the conservative (10th percentile) projections place saturation for most benchmarks before 2030.

2.3 Historical Context on Estimating AI Progress

Our projections should be interpreted against a recent pattern of systematic underestimation of AI progress [Kučinskas et al., 2025]. This contrasts with the earlier history of AI, with periods of excessive optimism about symbolic AI and expert systems, followed by "AI winters" in the 1970s and late 1980s. However, the current wave of deep learning progress, beginning around 2012, has consistently exceeded expectations. Since 2020, milestones in language understanding, mathematical reasoning, and code generation have arrived well ahead of expert forecasts [Steinhardt, 2022]. Expert forecasts have consistently predicted capability milestones further in the future than they actually occurred. For example, AI achieving gold-medal performance at the International Mathematical Olympiad was forecast for around 2030 by both domain and non-domain experts, with only 9% probability up to 2025, the year it was achieved [Kučinskas et al., 2025].

3 Implications and Call to Action

If AI systems achieve superhuman performance on most measurable cognitive tasks by 2030, several important implications follow.

3.1 For International Coordination

The goal of this Position Paper is not to prescribe specific policy actions but to ensure decision-makers recognize two points. First, GPAI capabilities are advancing rapidly, and closely monitoring frontier progress is important for making informed policy decisions. Second, regarding control and security-critical issues, the projected trajectory is not inevitable: it results from investment decisions [Sevilla et al., 2024], compute infrastructure build-out [Epoch AI, 2025f], and research priorities [Erdil and Besiroglu, 2023] that remain subject to collective choice. Given the projections presented in this article, the earlier multilateral discussions begin, the more options remain viable.

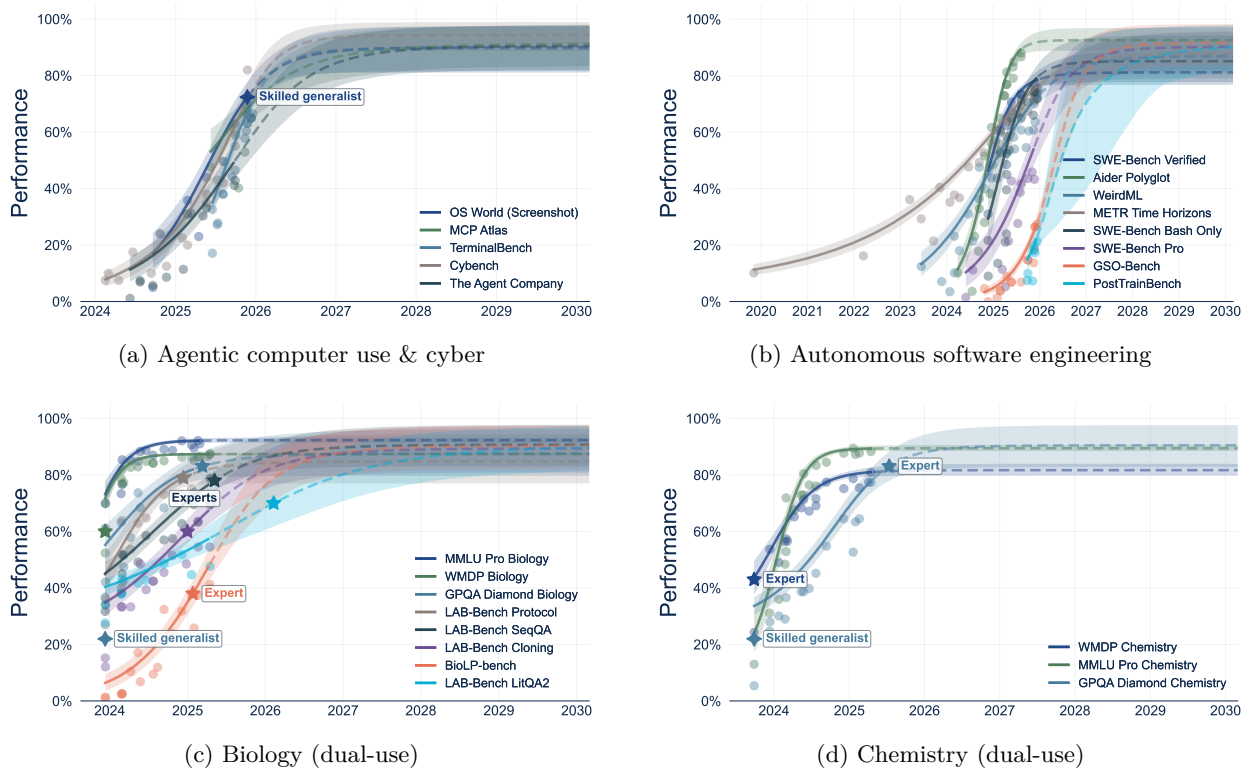
3.2 For AI Safety Research

The timeline for developing robust AI alignment techniques is short. Current approaches to alignment remain nascent: despite some progress, they still fail to ensure a high level of safety for current AI models, let alone scaling to systems significantly exceeding human capabilities [Amodei et al., 2016, Bengio et al., 2025]. Alignment is hard, which implies that by default these systems should not be expected to remain under human control [Maier et al., 2025, Ngo et al., 2025].

Given these short timelines, one cannot rely on any single research agenda succeeding. We recommend a portfolio approach:

- **Near-term agendas**: Prioritize research with potential payoffs within 3-5 years, such as technical governance agendas like robust verification mechanisms [Wasil et al., 2024, Scher and Thiergart, 2024].
- **Alternative architectures**: Invest in fundamentally different approaches (safe-by-design architectures, formal verification) to hedge against alignment failure in current paradigms [Dalrymple et al., 2024].
- **Differential technology development**: Prioritize research that advances safety and defensive capabilities over risk-increasing ones [Sandbrink et al., 2022].

We strongly underscore, however, that safety research alone cannot guarantee good outcomes if capability development continues regardless of safety progress.



(a) Agentic computer use & cyber

(b) Autonomous software engineering

(c) Biology (dual-use)

(d) Chemistry (dual-use)

Figure 4: **Security-critical capability trajectories.** Benchmarks with direct safety implications show rapid progress, with most projected to saturate by 2027-2028. Starred markers indicate human expert baselines where available. On OS World, unfamiliar users achieve 72% [Xie et al., 2024]; on biology benchmarks, PhD-level experts range from 38% (BioLP-bench) to 83% (GPQA Diamond Biology) [Dev et al., 2025, Rein et al., 2023]. Graphical conventions are similar to Figure 1.

3.3 For Evaluation Practices

Forecasts, such as those presented in this article, can only be as reliable as the data upon which they are based, and as long as the underlying dynamics driving the pace of progress remain unaltered. We detail below how these two sources of uncertainty could benefit from further research attention from the benchmarking and evaluation communities.

Benchmark validity. Do benchmark improvements reflect genuine capability gains, or artifacts of optimization, contamination, or narrow task-specific learning? This concern is not novel; prior work has called for harder benchmarks targeting real-world tasks, long-horizon planning, and dynamic evaluation protocols [Mazeika et al., 2025, White et al., 2025]. Our analysis already includes recently developed benchmarks designed to resist gaming, but connecting these scores to real-world impact remains understudied. Risk modeling efforts by Touzet et al. [2025], Barrett et al. [2025], Murray et al. [2025] to elicit expert knowledge in order to link capability metrics to potential harms represent a promising direction.

AI R&D acceleration. To what extent can AI systems automate AI research itself, potentially accelerating the pace of progress beyond current trends? Benchmarks measuring research capabilities (hypothesis generation, experiment design, code optimization) remain at an early stage [Kwa et al., 2025, Ihle, 2025]. We encourage the development of realistic evaluations for AI R&D capabilities, as these may provide early warning of acceleration dynamics and improve medium-horizon projections.

4 Alternative Views

We present credible counterarguments to our position and respond to each.

4.1 Benchmarks Can Be Gamed

Objection: Benchmark progress may reflect contamination (training on test data), reward-hacking on poorly designed tasks, and/or benchmark optimization (prompt engineering, finetuning on similar tasks, tuning scaffolds to evaluation formats) by AI companies incentivized to showcase impressive results with

each new model release [Robison, 2025]. Contamination in particular is a well-documented concern: Balloccu et al. [2024] show that several widely-used benchmarks have appeared in training corpora. If contamination and benchmark hacking inflate scores, especially for older benchmarks with longer exposure to training pipelines, then our hierarchical model could propagate these inflated trajectories to newer, less contaminated benchmarks.

Response: We acknowledge these concerns but note several mitigating factors: (1) The saturation trends presented in this article are consistent across 63 benchmarks from independent sources (Epoch AI, Scale AI, RAND). (2) These benchmarks are either run internally by these organizations or sourced from carefully selected benchmark developers [Epoch AI, 2024a] to limit contamination. (3) Many benchmarks saturate below 100%, which is consistent with labeling errors but inconsistent with wholesale answer memorization. (4) Newer benchmarks designed specifically to resist gaming (ARC-AGI, SWE-Bench Verified, FrontierMath) show similar trajectory patterns [Chollet, 2019, Epoch AI, 2024b, 2025c]. (5) Some benchmarks now evaluate models on tasks released after model training cutoffs, ruling out direct contamination [White et al., 2025]. While no benchmark is immune to all criticism, the convergent pattern across diverse, independently maintained evaluations provides evidence beyond any single benchmark. Furthermore, fitting each benchmark independently (without hierarchical pooling) yields similar conclusions (Supp. Table 2, $\geq 93\%$ of saturated benchmarks for all 8 model variants) with worse cross-validation fit (Supp. Table 1), suggesting that cross-benchmark information sharing does not artificially inflate projections for newer benchmarks.

4.2 Progress May Plateau

Objection: Scaling may hit diminishing returns. Data constraints, compute costs, or algorithmic limitations could slow progress before current or future benchmarks saturate. Concerns about a "data wall" (exhaustion of high-quality training data), plateauing returns from simply increasing compute (either for pre-training, post-training or inference scaling, see Ord [2025]), and the rising costs of frontier training runs all suggest that the current pace of improvement may not be sustainable. Each new scaling paradigm may buy time but itself face diminishing returns.

Response: This is possible, and our projections carry substantial uncertainty. However, (1) predicted slowdowns in the past decade have repeatedly failed to materialize, (2) new scaling paradigms (reasoning models, inference-time compute) continue to unlock progress, as the accelerating absolute investment in AI compute and research currently compensates diminishing returns, and (3) looking at each potential bottleneck individually, scaling seems to be on track to continue for at least the end of the decade [Sevilla et al., 2024].

4.3 Benchmarks Do Not Measure General Intelligence

Objection: Benchmark saturation does not imply human-level intelligence. Models may achieve high scores through pattern matching, heuristics, or memorization rather than genuine understanding. This is a longstanding concern in AI evaluation [Chollet, 2019]: high benchmark scores may reflect narrow task-specific optimization rather than the flexible, transferable reasoning that characterizes human cognition. A model that solves FrontierMath problems may do so through learned solution templates rather than on-the-fly reasoning, as the impressive math skills of recent AI models seem to rely more on their extensive knowledge of the literature and of known techniques than on reasoning depth. Correspondingly, AI progress is "jagged", superhuman on some tasks yet subhuman on others. This heterogeneity could persist, preventing AGI-like systems that could reliably replace full human jobs or achieve their goals autonomously far outside of their training distribution.

Response: Benchmark performance is indeed an imperfect proxy for general intelligence; this is why we frame our position around "measurable cognitive tasks" rather than AGI. However, (1) many recent benchmarks specifically target capabilities thought to require flexible and general reasoning [ARC Prize, 2025a, Wang et al., 2025] or the ability to handle real-world under-specified tasks [Mazeika et al., 2025], (2) as gaps are progressively closing, the breadth of saturation across diverse tasks becomes more and more difficult to explain by narrow optimization alone, and (3) if AI reaches superhuman performance on AI R&D itself, even if it is short of general intelligence, remaining gaps may close rapidly through recursive improvement.

Lastly, from a practical standpoint, systems that exceed human performance on most measurable cognitive tasks could start to have irreversible global impacts regardless of whether they possess true general intelligence. Nevertheless, we acknowledge that this is a crux and a significant source of uncertainty in predicting the consequences of benchmark saturation.

4.4 Benchmark Saturation And Superhuman Scores Do Not Entail Irreversible Global Impacts

Objection: First, benchmark saturation reflects an intrinsic property of their construction based on a fixed set of tasks. So even if AI saturates all existing benchmarks and exceeds expert performance,

this may not translate to real-world transformative capabilities. One reason is that as our analysis is limited to cognitive tasks that can be evaluated by online benchmarks, it does not cover embodied capabilities, or social interactions, which may remain significant bottlenecks for real-world impact even if cognitive benchmarks are saturated. AI systems have matched or exceeded radiologist performance on medical imaging benchmarks for years, yet the demand for radiologists is greater than ever. Second, even with these wider capabilities unlocked, numerous frictions and regulatory constraints will likely delay the deployment of AI systems in society. Finally, humans have always created tools that exceed their own capabilities in specific domains. Computers can process information faster than any human, and planes travel at superhuman speeds. AI is the latest such technology, and its development warrants proportionate regulation (as for the aviation industry), not alarmist framing.

Response: We agree that benchmark saturation is a necessary but not sufficient condition for irreversible real-world impact, and address each argument in turn. On the significance of benchmark saturation: the benchmarks in this article do have ceilings, so their eventual saturation is indeed expected. However, the *speed* of saturation is informative. Several benchmarks in our set were designed to resist rapid saturation (FrontierMath, Humanity's Last Exam, the ARC-AGI series). That these benchmarks could be approaching saturation before 2028, with domain experts and even committees of experts failing to beat the latest AI models, would have been considered a very optimistic prediction a few years ago. Furthermore, the benchmarks most relevant to our security argument (cybersecurity, AI R&D automation, CBRN knowledge) do not require embodiment to translate into real-world impact. On deployment frictions: these are real for many beneficial applications (regulatory approval, barriers to adoption, etc.), but they are greatly reduced both for AI companies deploying AI R&D capabilities internally and for adversarial actors. Our security concerns primarily target these faster pathways. On the tools analogy: the comparison to calculators or planes understates three properties that frontier AI systems are currently acquiring and that have not been combined before. These are (a) general performance across dozens of cognitive domains, (b) autonomous agentic operation, and (c) the potential to accelerate its own development. We believe that this combination, paired with the current absence of reliable control methods, distinguishes frontier AI from prior technologies.

More broadly, the aim of our paper is to provide evidence on the size of the window for preventive action. The full consequences of saturating cybersecurity, CBRN, AI R&D and remote work benchmarks are uncertain, but we note that this progress is for the most part irreversible (more so when considering the fast following of cheaper open source models). So waiting for the gap between benchmarks and real-world impact to resolve means allowing the window for effective collective action to close.

4.5 What Would Change Our Mind

Our position would be substantially weakened by:

- Sustained plateau in benchmark progress (> 2 years of stagnation across multiple hard benchmarks);
- Evidence that current architectures face fundamental limits on specific capability dimensions, especially on capabilities necessary to expert-level AI R&D automation;
- Demonstration that benchmark performance systematically diverges from real-world task performance, e.g. a continued plateau of the Remote Labor Index [Mazeika et al., 2025] or the absence of a noticeable impact of AI automation on the US economic growth by 2030.

5 Limitations

Our analysis has several limitations:

Extrapolation uncertainty. All forecasting involves extrapolation. Our Bayesian approach quantifies some sources of uncertainty, but it does not explicitly integrate the many known and unknown factors that may disrupt future trajectories in either direction (scaling bottlenecks, algorithmic breakthroughs, AI R&D automation, geopolitical conflicts, international agreements, etc.).

Elicitation gap. Benchmark scores underestimate latent model capabilities due to suboptimal prompting, scaffolding, evaluation conditions, or even sandbagging (i.e., strategic underperformance, see Weij et al. [2025]). Our skewed likelihood model partially addresses this issue by inferring a latent frontier capability mostly above the observed top-3 scores. Nonetheless, if substantial capabilities remain hidden due to inadequate evaluation protocols or secrecy in AI development, our projections would be conservative. The true trajectory could thus be faster than our estimates suggest.

Modeling extensions. Several extensions could improve forecast accuracy: (1) incorporating compute scaling as a predictor, linking benchmark progress to training resources [Kaplan et al., 2020, Ruan et al., 2024]; (2) modeling latent capability dimensions that manifest across multiple benchmarks [Burnell et al., 2023, Kipnis et al., 2025, Ho et al., 2025, Polo et al.,

2025, Pimpale et al., 2025]; (3) integrating information about model architectures and training approaches. We chose a simpler model to maintain interpretability, but richer models may become feasible as more data accumulates.

Scope of analysis. Our analysis is limited to cognitive tasks that can be evaluated through benchmarks. It does not cover non-cognitive capabilities such as embodiment, physical manipulation, or social interactions, which may be critical bottlenecks for certain real-world impacts even if cognitive benchmarks are saturated. The relationship between benchmark saturation and transformative real-world capability remains an open question.

Benchmark evolution. Our analysis covers only existing benchmarks, and harder task sets are continuously being developed. However, in many domains, frontier models already match or exceed expert human performance, raising questions about how much headroom remains. Designing "superhuman" evaluations requires either (1) tasks where ground truth is verifiable without human judgment (e.g., mathematical proofs, code execution), or (2) aggregating across many human experts [Phan et al., 2025]. Some benchmarks already approximate this standard by comparing AI to real-world human task completion. A potential next step to forecast AI progress beyond current benchmarks is to extrapolate higher-level properties of the evaluation tasks (e.g., human-equivalent time horizon, see Kwa et al. [2025], or task difficulty level, see Zhou et al. [2025]).

6 Conclusion

We have argued that current AI development trajectory leads to the saturation of most existing cognitive benchmarks by 2030, approaching or exceeding human expert performance across a broad range of measurable cognitive tasks. Given the state of our understanding of and control over these systems, the cost of inaction could be severe. This trajectory is not inevitable, as it results from choices that can still be influenced through coordinated action, but the window is closing rapidly.

We call on the ML community and institutional actors to engage with this possibility and act accordingly: differentially accelerating safety research, monitoring critical AI capabilities, and developing verification mechanisms as part of a global effort to provide security guarantees commensurate to the possibility of superhuman AIs in the coming years.

Reproducibility Statement

All benchmark data are publicly available from Epoch AI [Epoch AI, 2024a], Scale AI [Scale AI, 2025], and RAND [Dev et al., 2025]. Code for model fitting and visualization is available at `https://github.com/General-Purpose-AI-Policy-Lab/benchmark-forecasting`.

References

Epoch AI. AI Benchmarking Hub, November 2024a. URL `https://epoch.ai/benchmarks/use-this-data`.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. International AI Safety Report. Technical Report arXiv:2501.17805, arXiv, January 2025. URL `http://arxiv.org/abs/2501.17805`. arXiv:2501.17805 [cs].

AISI. Frontier AI Trends Report by The AI Security Institute (AISI). Technical report, AI Security Institute, 2025. URL `https://www.aisi.gov.uk/frontier-ai-trends-report`.

Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When Will AI Exceed Human Performance? Evidence from AI Experts, May 2018. URL `http://arxiv.org/abs/1705.08807`. arXiv:1705.08807 [cs].

Katja Grace, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, Jan Brauner, and Richard C. Korzekwa. Thousands of AI Authors on the Future of AI. *jair*, 84, October 2025. ISSN 1076-9757. doi: 10.1613/jair.1.19087. URL http://arxiv.org/abs/2401.02843. arXiv:2401.02843 [cs].

François Chollet. On the Measure of Intelligence, November 2019. URL http://arxiv.org/abs/1911.01547. arXiv:1911.01547 [cs].

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, September 2023. URL http://arxiv.org/abs/2307.15217. arXiv:2307.15217 [cs].

Antoine Maier, Aude Maier, and Tom David. Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization, October 2025. URL http://arxiv.org/abs/2510.02840. arXiv:2510.02840 [cs].

Richard Ngo, Lawrence Chan, and Sören Mindermann. The Alignment Problem from a Deep Learning Perspective, May 2025. URL http://arxiv.org/abs/2209.00626. arXiv:2209.00626 [cs].

Scale AI. Introducing VISTA: A Rubric-based Visual Task Assessment, 2025. URL https://scale.com/leaderboard/visual_language_understanding.

Sunishchal Dev, Charles Teague, Grant Ellison, Kyle Brady, Ying-Chiang Jeffrey Lee, Sarah L. Gebauer, Henry Alexander Bradley, Dawid Maciorowski, Bria Persaud, Jordan Despanie, Barbara Del Castello, Alyssa Worland, Michael Miller, Adrian Salas, Dave Nguyen, James Liu, Jason Johnson, Andrew Sloan, Will Stonehouse, Travis Merrill, Thomas Goode, Greg McKelvey, and Ella Guest. Toward Comprehensive Benchmarking of the Biological Knowledge of Frontier Large Language Models. Technical report, RAND Corporation, 2025. URL https://www.rand.org/pubs/research_reports/RRA3797-1.html.

A. C. Harvey. Time Series Forecasting Based on the Logistic Curve. *Journal of the Operational Research Society*, 35(7):641–646, July 1984. ISSN 0160-5682. doi: 10.1057/jors.1984.128. URL https://doi.org/10.1057/jors.1984.128. _eprint: https://doi.org/10.1057/jors.1984.128.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering, October 2022. URL http://arxiv.org/abs/2209.09513. arXiv:2209.09513 [cs].

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, March 2018. URL http://arxiv.org/abs/1803.05457. arXiv:1803.05457 [cs].

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, November 2021. URL http://arxiv.org/abs/2110.14168. arXiv:2110.14168 [cs].

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, November 2023. URL http://arxiv.org/abs/2311.12022. arXiv:2311.12022 [cs].

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

Afra Feyza Akyürek, Advait Gosai, Chen Bo Calvin Zhang, Vipul Gupta, Jaehwan Jeong, Anisha Gunjal, Tahseen Rabbani, Maria Mazzone, David Randolph, Mohammad Mahmoudi Meymand, Gurshaan Chattha, Paula Rodriguez, Diego Mares, Pavit Singh, Michael Liu, Subodh Chawla, Pete Cline, Lucy Ogaz, Ernesto Hernandez, Zihao Wang, Pavi Bhatter, Marcos Ayestaran, Bing Liu, and Yunzhong He. PRBench: Large-Scale Expert Rubrics for Evaluating High-Stakes Professional Reasoning, November 2025. URL http://arxiv.org/abs/2511.11562. arXiv:2511.11562 [cs] version: 1.

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. SimpleQA Verified: A Reliable Factuality Benchmark to Measure Parametric Knowledge, September 2025. URL http://arxiv.org/abs/2509.07968. arXiv:2509.07968 [cs].

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, Søren 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, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, 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, Martí 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, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, 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 Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H. Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, 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, J. P. 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, Jérémy Andréoletti, Raghav Singhal, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khánh, 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, Biró Bálint, Eve J. Y. Lo, Jiaqi Wang, Maria Inês 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 Ciobâcă, 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 Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, 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, Yiğit Yalın, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Boscá, 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 Häggström, 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 Hernández-Cámara, Emanuele Rodolà, 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 José Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Duc 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, Nico-

las 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 Łucki, 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 Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás 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 Zubić, 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, Vilém 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, Mickaël Noyé, Michał Perełkiewicz, 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, Dániel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Ginés, 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 Lù, 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 Briański, 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 Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E. Zhang, Zhun Wang, Gaël 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, Károly Zsolnai-Fehér, 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. Gonçalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Yücel, 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, Summer Yue, Alexandr Wang, and Dan Hendrycks. Humanity's Last Exam, September 2025. URL http://arxiv.org/abs/2501.14249. arXiv:2501.14249 [cs].

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A New Benchmark for Natural Language Understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proc. of ACL*, pages 4885–4901, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

Philip and Hemang. SimpleBench: The Text Benchmark in which Unspecialized Human Performance Exceeds that of Current Frontier Models, 2024. URL https://simple-bench.com.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Lukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. BALROG: Benchmarking Agentic LLM and VLM Reasoning On Games, April 2025. URL http://arxiv.org/abs/2411.13543. arXiv:2411.13543 [cs].

Epoch AI. Chess Puzzles, December 2025a. URL https://epoch.ai/benchmarks/chess-puzzles.

Clinton J. Wang, Dean Lee, Cristina Menghini, Johannes Mols, Jack Doughty, Adam Khoja, Jayson Lynch, Sean Hendryx, Summer Yue, and Dan Hendrycks. EnigmaEval: A Benchmark of Long Multimodal Reasoning Challenges, February 2025. URL http://arxiv.org/abs/2502.08859. arXiv:2502.08859 [cs].

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, November 2021. URL http://arxiv.org/abs/2103.03874. arXiv:2103.03874 [cs].

Epoch AI. OTIS Mock AIME 2024-2025, 2025b. URL https://epoch.ai/benchmarks/otis-mock-aime-2024-2025.

Epoch AI. FrontierMath, 2025c. URL https://epoch.ai/benchmarks/frontiermath.

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, April 2025. URL http://arxiv.org/abs/2406.19314. arXiv:2406.19314 [cs].

ARC Prize. ARC-AGI-1, 2019. URL https://arcprize.org/arc-agi/1/.

ARC Prize. ARC-AGI-2, 2025a. URL https://arcprize.org/arc-agi/2/.

ARC Prize. ARC-AGI-3, 2025b. URL https://arcprize.org/arc-agi/3/.

Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwag, Shivam Singhal, Yury Orlovskiy, Steven Basart, Manasi Sharma, Denis Peskoff, Elaine Lau, Jaehyuk Lim, Lachlan Carroll, Alice Blair, Vinaya Sivakumar, Sumana Basu, Brad Kenstler, Yuntao Ma, Julian Michael, Xiaoke Li, Oliver Ingebretsen, Aditya Mehta, Jean Mottola, John Teichmann, Kevin Yu, Zaina Shaik, Adam Khoja, Richard Ren, Jason Hausenloy, Long Phan, Ye Htet, Ankit Aich, Tahseen Rabbani, Vivswan Shah, Andriy Novykov, Felix Binder, Kirill Chugunov, Luis Ramirez, Matias Geralnik, Hernán Mesura, Dean Lee, Ed-Yeremai Hernandez Cardona, Annette Diamond, Summer Yue, Alexandr Wang, Bing Liu, Ernesto Hernandez, and Dan Hendrycks. Remote Labor Index: Measuring AI Automation of Remote Work, October 2025. URL http://arxiv.org/abs/2510.26787. arXiv:2510.26787 [cs].

Bertie Vidgen, Austin Mann, Abby Fennelly, John Wright Stanly, Lucas Rothman, Marco Burstein, Julien Benchek, David Ostrofsky, Anirudh Ravichandran, Debnil Sur, Neel Venugopal, Alannah Hsia, Isaac Robinson, Calix Huang, Olivia Varones, Daniyal Khan, Michael Haines,

Austin Bridges, Jesse Boyle, Koby Twist, Zach Richards, Chirag Mahapatra, Brendan Foody, and Osvald Nitski. APEX-Agents, February 2026. URL http://arxiv.org/abs/2601.14242. arXiv:2601.14242 [cs].

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks, October 2025. URL http://arxiv.org/abs/2510.04374. arXiv:2510.04374 [cs].

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, Junhong Shen, Guanghao Ye, Haowei Lin, Jason Poulos, Maoyu Wang, Marianna Nezhurina, Jenia Jitsev, Di Lu, Orfeas Menis Mastromichalakis, Zhiwei Xu, Zizhao Chen, Yue Liu, Robert Zhang, Leon Liangyu Chen, Anurag Kashyap, Jan-Lucas Uslu, Jeffrey Li, Jianbo Wu, Minghao Yan, Song Bian, Vedang Sharma, Ke Sun, Steven Dillmann, Akshay Anand, Andrew Lanpouthakoun, Bardia Koopah, Changran Hu, Etash Guha, Gabriel H. S. Dreiman, Jiacheng Zhu, Karl Krauth, Li Zhong, Niklas Muennighoff, Robert Amanfu, Shangyin Tan, Shreyas Pimpalgaonkar, Tushar Aggarwal, Xiangning Lin, Xin Lan, Xuandong Zhao, Yiqing Liang, Yuanli Wang, Zilong Wang, Changzhi Zhou, David Heineman, Hange Liu, Harsh Trivedi, John Yang, Junhong Lin, Manish Shetty, Michael Yang, Nabil Omi, Negin Raoof, Shanda Li, Terry Yue Zhuo, Wuwei Lin, Yiwei Dai, Yuxin Wang, Wenhao Chai, Shang Zhou, Dariush Wahdany, Ziyu She, Jiaming Hu, Zhikang Dong, Yuxuan Zhu, Sasha Cui, Ahson Saiyed, Arinbjörn Kolbeinsson, Jesse Hu, Christopher Michael Rytting, Ryan Marten, Yixin Wang, Alex Dimakis, Andy Konwinski, and Ludwig Schmidt. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces, January 2026. URL http://arxiv.org/abs/2601.11868. arXiv:2601.11868 [cs].

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Andy K. Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin W. Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Jasper, Pura Peetathawatchai, Ari Glenn, Vikram Sivashankar, Daniel Zamoshchin, Leo Glikbarg, Derek Askaryar, Mike Yang, Teddy Zhang, Rishi Alluri, Nathan Tran, Rinnara Sangpisit, Polycarpos Yiorkadjis, Kenny Osele, Gautham Raghupathi, Dan Boneh, Daniel E. Ho, and Percy Liang. Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models, April 2025. URL http://arxiv.org/abs/2408.08926. arXiv:2408.08926 [cs].

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks, September 2025. URL http://arxiv.org/abs/2412.14161. arXiv:2412.14161 [cs].

Chaithanya Bandi, Ben Hertzberg, Geobio Boo, Tejas Polakam, Jeff Da, Sami Hassaan, Manasi Sharma, Andrew Park, Ernesto Hernandez, Dan Rambado, Ivan Salazar, Chetan Rane, Ben Levin, Brad Kenstler, and Bing Liu. MCP-Atlas: A Large-Scale Benchmark for Tool-Use Competency with Real MCP Servers, December 2025.

aider. Aider LLM Leaderboards, 2024. URL https://aider.chat/docs/leaderboards/.

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, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI Ability to Complete Long Tasks, March 2025. URL http://arxiv.org/abs/2503.14499. arXiv:2503.14499 [cs].

Håvard Tveit Ihle. WeirdML, 2025. URL https://htihle.github.io/weirdml.html.

Epoch AI. SWE-bench Verified, 2024b. URL https://epoch.ai/benchmarks/swe-bench-verified.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R.

Narasimhan. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *Proc. of ICLR*. OpenReview.net, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, Karmini Sampath, Maya Krishnan, Srivatsa Kundurthy, Sean Hendryx, Zifan Wang, Vijay Bharadwaj, Jeff Holm, Raja Aluri, Chen Bo Calvin Zhang, Noah Jacobson, Bing Liu, and Brad Kenstler. SWE-Bench Pro: Can AI Agents Solve Long-Horizon Software Engineering Tasks?, November 2025. URL http://arxiv.org/abs/2509.16941. arXiv:2509.16941 [cs].

Manish Shetty, Naman Jain, Jinjian Liu, Vijay Kethanaboyina, Koushik Sen, and Ion Stoica. GSO: Challenging Software Optimization Tasks for Evaluating SWE-Agents, October 2025. URL http://arxiv.org/abs/2505.23671. arXiv:2505.23671 [cs].

Ben Rank, Hardik Bhatnagar, Ameya Prabhu, Shira Eisenberg, Karina Nguyen, Matthias Bethge, and Maksym Andriushchenko. PostTrainBench: Can LLM Agents Automate LLM Post-Training?, March 2026. URL http://arxiv.org/abs/2603.08640. arXiv:2603.08640 [cs].

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning, May 2024a. URL http://arxiv.org/abs/2403.03218. arXiv:2403.03218 [cs].

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, July 2024. URL http://arxiv.org/abs/2407.10362. arXiv:2407.10362 [cs].

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering, September 2018. URL http://arxiv.org/abs/1809.02789. arXiv:1809.02789 [cs].

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proc. of ACL*, pages 4791–4800, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439, April 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6399.

Lech Mazur. lechmazur/writing, January 2026. URL https://github.com/lechmazur/writing. original-date: 2025-01-05T08:48:04Z.

Epoch AI. Fiction.liveBench, February 2025d. URL https://epoch.ai/benchmarks/fictionlivebench.

Rakshith S. Srinivasa, Zora Che, Chen Bo Calvin Zhang, Diego Mares, Ernesto Hernandez, Jayeon Park, Dean Lee, Guillermo Mangialardi, Charmaine Ng, Ed-Yeremai Hernandez Cardona, Anisha Gunjal, Yunzhong He, Bing Liu, and Chen Xing. TutorBench: A Benchmark To Assess Tutoring Capabilities Of Large Language Models, October 2025. URL http://arxiv.org/abs/2510.02663. arXiv:2510.02663 [cs].

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs, March 2025. URL http://arxiv.org/abs/2501.17399. arXiv:2501.17399 [cs].

Alexander R. Fabbri, Diego Mares, Jorge Flores, Meher Mankikar, Ernesto Hernandez, Dean Lee, Bing Liu, and Chen Xing. MultiNRC: A Challenging and Native Multilingual Reasoning Evaluation Benchmark for LLMs, July 2025. URL http://arxiv.org/abs/2507.17476. arXiv:2507.17476 [cs].

ccmdi. ccmdi/geobench, January 2026. URL https://github.com/ccmdi/geobench. original-date: 2025-03-22T13:27:48Z.

Epoch AI. CadEval, 2025e. URL https://epoch.ai/benchmarks/cad-eval.

Chase Brower. Visual Physics Comprehension Test, 2025. URL https://cbrower.dev/vpct.

Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, and Yunzhong He. Audio MultiChallenge: A Multi-Turn Evaluation of Spoken Dialogue Systems on Natural Human Interaction, December 2025. URL http://arxiv.org/abs/2512.14865. arXiv:2512.14865 [cs].

Xingang Guo, Utkarsh Tyagi, Advait Gosai, Paula Vergara, Jayeon Park, Ernesto Gabriel Hernández Montoya, Chen Bo Calvin Zhang, Bin Hu, Yunzhong He, Bing Liu, and Rakshith Sharma Srinivasa. Beyond Seeing: Evaluating Multimodal LLMs on Tool-Enabled Image Perception, Transformation, and Reasoning, October 2025. URL http://arxiv.org/abs/2510.12712. arXiv:2510.12712 [cs].

Simas Kučinskas, Josh Rosenberg, Rebecca Ceppas de Castro, Zach Jacobs, Jordan Canedy, Philip E Tetlock, and Ezra Karger. Assessing Near-Term Accuracy in the Existential Risk Persuasion Tournament. Technical report, Forecasting Research Institute, September 2025. URL https://forecastingresearch.org/near-term-xpt-accuracy.

Jacob Steinhardt. AI Forecasting: One Year In, July 2022. URL https://bounded-regret.ghost.io/ai-forecasting-one-year-in/.

Jaime Sevilla, Tamay Besiroglu, Ben Cottier, Josh You, Edu Roldán, Pablo Villalobos, and Ege Erdil. Can AI scaling continue through 2030? Technical report, Epoch AI, 2024. URL https://epoch.ai/blog/can-ai-scaling-continue-through-2030.

Epoch AI. Data on Frontier AI Data Centers, September 2025f. URL https://epoch.ai/data/data-centers.

Ege Erdil and Tamay Besiroglu. Algorithmic progress in computer vision, August 2023. URL http://arxiv.org/abs/2212.05153. arXiv:2212.05153 [cs].

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety, July 2016. URL http://arxiv.org/abs/1606.06565. arXiv:1606.06565 [cs].

Akash R. Wasil, Peter Barnett, Michael Gerovitch, Roman Hauksson, Tom Reed, and Jack William Miller. Governing dual-use technologies: Case studies of international security agreements and lessons for AI governance, September 2024. URL http://arxiv.org/abs/2409.02779. arXiv:2409.02779 [cs] version: 1.

Aaron Scher and Lisa Thiergart. Mechanisms to Verify International Agreements About AI Development. Technical report, Machine Intelligence Research Institute, November 2024. URL https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development.

David "davidad" Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems, July 2024. URL http://arxiv.org/abs/2405.06624. arXiv:2405.06624 [cs].

Jonas Sandbrink, Hamish Hobbs, Jacob Swett, Allan Dafoe, and Anders Sandberg. Differential technology development: An innovation governance consideration for navigating technology risks, September 2022. URL https://papers.ssrn.com/abstract=4213670.

Chloé Touzet, Henry Papadatos, Malcolm Murray, Otter Quarks, Steve Barrett, Alejandro Tlaie Boria, Elija Perrier, Matthew Smith, and Siméon Campos. The Role of Risk Modeling in Advanced AI Risk Management, December 2025. URL http://arxiv.org/abs/2512.08723. arXiv:2512.08723 [cs].

Steve Barrett, Malcolm Murray, Otter Quarks, Matthew Smith, Jakub Kryś, Siméon Campos, Alejandro Tlaie Boria, Chloé Touzet, Sevan Hayrapet, Fred Heiding, Omer Nevo, Adam Swanda, Jair Aguirre, Asher Brass Gershovich, Eric Clay, Ryan Fetterman, Mario Fritz, Marc Juarez, Vasilios Mavroudis, and Henry Papadatos. Toward Quantitative Modeling of Cybersecurity Risks Due to AI Misuse, December 2025. URL http://arxiv.org/abs/2512.08864. arXiv:2512.08864 [cs].

Malcolm Murray, Steve Barrett, Henry Papadatos, Otter Quarks, Matt Smith, Alejandro Tlaie Boria, Chloé Touzet, and Siméon Campos. A Methodology for Quantitative AI Risk Modeling, December 2025. URL http://arxiv.org/abs/2512.08844. arXiv:2512.08844 [cs].

Kylie Robison. Meta got caught gaming AI benchmarks, April 2025. URL https://www.theverge.com/meta/645012/meta-llama-4-maverick-benchmarks-gaming.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs, February 2024. URL http://arxiv.org/abs/2402.03927. arXiv:2402.03927 [cs].

Toby Ord. Inference Scaling Reshapes AI Governance, February 2025. URL http://arxiv.org/abs/2503.05705. arXiv:2503.05705 [cs].

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI Sandbagging: Language Models can Strategically Underperform on Evaluations, February 2025. URL http://arxiv.org/abs/2406.07358. arXiv:2406.07358 [cs].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs].

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational Scaling Laws and the Predictability of Language Model Performance, October 2024. URL http://arxiv.org/abs/2405.10938. arXiv:2405.10938 [cs].

Ryan Burnell, Han Hao, Andrew R. A. Conway, and Jose Hernandez Orallo. Revealing the structure of language model capabilities, June 2023. URL http://arxiv.org/abs/2306.10062. arXiv:2306.10062 [cs].

Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench – A Sparse Benchmark of Reasoning and Knowledge in Large Language Models, February 2025. URL http://arxiv.org/abs/2407.12844. arXiv:2407.12844 [cs].

Anson Ho, Jean-Stanislas Denain, David Atanasov, Samuel Albanie, and Rohin Shah. A Rosetta Stone for AI Benchmarks, November 2025. URL http://arxiv.org/abs/2512.00193. arXiv:2512.00193 [cs].

Felipe Maia Polo, Seamus Somerstep, Leshem Choshen, Yuekai Sun, and Mikhail Yurochkin. Sloth: scaling laws for LLM skills to predict multi-benchmark performance across families, December 2025. URL http://arxiv.org/abs/2412.06540. arXiv:2412.06540 [cs].

Govind Pimpale, Axel Højmark, Jérémy Scheurer, and Marius Hobbhahn. Forecasting Frontier Language Model Agent Capabilities, March 2025. URL http://arxiv.org/abs/2502.15850. arXiv:2502.15850 [cs].

Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M. Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E. Prunty, Zongqian Li, Pablo Sánchez-García, Kexin Jiang Chen, Pablo A. M. Casares, Jiyun Zu, John Burden, Behzad Mehrbakhsh, David Stillwell, Manuel Cebrian, Jindong Wang, Peter Henderson, Sherry Tongshuang Wu, Patrick C. Kyllonen, Lucy Cheke, Xing Xie, and José Hernández-Orallo. General Scales Unlock AI Evaluation with Explanatory and Predictive Power, March 2025. URL http://arxiv.org/abs/2503.06378. arXiv:2503.06378 [cs].

Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J. Fonnesbeck, Maxim Kochurov, Ravin Kumar, Junpeng Lao, Christian C. Luhmann, Osvaldo A. Martin, Michael Osthege, Ricardo Vieira, Thomas Wiecki, and Robert Zinkov. PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Comput. Sci.*, 9:e1516, September 2023. ISSN 2376-5992. doi: 10.7717/peerj-cs.1516. URL https://peerj.com/articles/cs-1516.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput*, 27(5):1413–1432, September 2017. ISSN 1573-1375. doi: 10.1007/s11222-016-9696-4. URL https://doi.org/10.1007/s11222-016-9696-4.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

OpenAI. Learning to Reason with LLMs, September 2024. URL https://openai.com/index/learning-to-reason-with-llms.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-Plus: A Comprehensive Benchmark for Evaluating the Robustness of LLMs as Mathematical Problem Solvers, July 2024b. URL http://arxiv.org/abs/2402.19255. arXiv:2402.19255 [cs].

Evan Chen. OTIS Mock AIME 2025 Report. Technical report, Evan Chen Website, 2025. URL https://web.evanchen.cc/exams/sols-OTIS-Mock-AIME-2025.pdf.

A Methodological Details

A.1 Growth Curve Specification

Let $y_i(t)$ denote the observed score for benchmark i at time t (days since first observation). We model the latent frontier performance as a shifted sigmoid:

$$\mu_i(t) = \ell_i + (L_i - \ell_i) \cdot \sigma_i(t),$$

where $\ell_i \in [0, 1]$ is the benchmark-specific lower bound (random-chance performance, manually gathered per benchmark, set to 0 when unknown) and $L_i \in [\ell_i, 1]$ is the upper asymptote (inferred).

Harvey function. The Harvey curve generalizes the logistic with a shape parameter $\alpha_i > 1$:

$$\sigma_i^{\text{Harvey}}(t) = [1 - (1 - \alpha_i) \exp(-k_i(t - \tau_i))]^{\frac{1}{1-\alpha_i}},$$

where $k_i > 0$ is the growth rate and τ_i is the inflection time. The parameter α_i controls asymmetry: larger values produce more gradual accelerations and faster saturation. When $\alpha_i = 2$, the Harvey function reduces to the standard logistic.

Logistic function. For comparison, we also fit the standard logistic:

$$\sigma_i^{\text{Logistic}}(t) = \frac{1}{1 + \exp(-k_i(t - \tau_i))}.$$

A.2 Observation Model

Observations follow a skew-normal distribution with heteroskedastic noise:

$$y_i(t) \sim \text{SkewNormal}(\mu_i(t), \xi_i(t), s_i),$$

where $s_i \leq 0$ is a skewness parameter allowing scores to fall predominantly below the latent curve.

The noise scale $\xi_i(t)$ is heteroskedastic and Beta-shaped over $[\ell_i, L_i]$:

$$\xi_i(t) = \xi_0 + \xi_i^{\text{base}} \cdot \frac{\sqrt{(\mu_i(t) - \ell_i)(L_i - \mu_i(t))}}{(L_i - \ell_i)/2},$$

with $\xi_0 = 0.01$ fixed. This yields maximal variance near the inflection point and minimal variance near the bounds.

A.3 Hierarchical Structure

The joint model places shared hyperpriors across benchmarks:

Upper asymptote. $L_i = L_{\min} + (1 - L_{\min}) \cdot L_i^{\text{raw}}$, with $L_{\min} = 0.75$ and $L_i^{\text{raw}} \sim \text{Beta}(\mu_L, \sigma_L)$, and hyperpriors $\mu_L \sim \text{Beta}\left(\text{mean} = \frac{0.96 - L_{\min}}{1 - L_{\min}}, \text{sd} = \frac{0.02}{1 - L_{\min}}\right)$ and $\sigma_L \sim \text{Half-}\mathcal{N}\left(\text{sd} = \frac{0.02}{1 - L_{\min}}\right)$.

Growth rate. $k_i \sim \mathcal{G}(k_\mu, k_\sigma)$, with $k_\mu \sim \mathcal{G}(\text{mean} = 0.005, \text{sd} = 0.002)$ and $k_\sigma \sim \text{Half-}\mathcal{N}(\text{sd} = 0.005)$.

Inflection time. $\tau_i \sim \text{Gumbel}(\hat{\tau}_i, \beta)$, where $\hat{\tau}_i$ is the empirical midpoint of benchmark i's observed time range and $\beta = 730$ days (2 years). The right-skewed Gumbel prior reflects that for unsaturated benchmarks, the inflection point likely lies beyond observed data.

Noise. $\xi_i^{\text{base}} \sim \mathcal{G}(\xi_\mu^{\text{base}}, \xi_\sigma^{\text{base}})$, with hyperpriors $\xi_\mu^{\text{base}} \sim \mathcal{G}(\text{mean} = 0.05 + \frac{N}{50}, \text{sd} = 0.02)$ and $\xi_\sigma^{\text{base}} \sim \text{Half-}\mathcal{N}(\text{sd} = 0.05)$, where N corresponds to the top-N frontier of scores considered ($N = 3$ in this paper).

Skewness. $s_i \sim \text{Truncated-}\mathcal{N}(s_\mu, s_\sigma, -\infty, 0)$, with hyperpriors $s_\mu \sim \mathcal{N}(\text{mean} = -2 - \frac{N}{2}, \text{sd} = 0.5)$ and $s_\sigma \sim \text{Half-}\mathcal{N}(\text{sd} = 1.0)$.

Harvey shape. $\alpha_i = 1 + \alpha_i^{\text{raw}}$, with $\alpha_i^{\text{raw}} \sim \mathcal{G}(\alpha_\mu^{\text{raw}}, \alpha_\sigma^{\text{raw}})$, ensuring $\alpha_i > 1$, and hyperpriors $\alpha_\mu^{\text{raw}} \sim \mathcal{G}(\text{mean} = 1.5, \text{sd} = 0.5)$ and $\alpha_\sigma^{\text{raw}} \sim \text{Half-}\mathcal{N}(\text{sd} = 0.5)$.

A.4 Inference

We track the top-3 frontier scores per benchmark at each point in time to reduce noise from suboptimal evaluations. Model fitting uses MCMC via PyMC [Abril-Pla et al., 2023] with the NUTS sampler (2000 posterior samples, 1000 warmup iterations, 4 chains, target acceptance 0.9). Convergence is assessed via $\hat{R}$ diagnostics and effective sample size.

B Sensitivity Analysis

We assess the robustness of our main finding to all modeling choices by evaluating a full grid of 8 model variants: sigmoid family (Harvey vs. logistic) × hierarchical structure (joint vs. independent) × likelihood (skew-normal vs. normal). Independent fits use identical model structure but estimate all parameters separately for each benchmark without shared hyperpriors. Each variant is evaluated on model comparison via LOO-ELPD and saturation projections.

B.1 Model Comparison (LOO-ELPD)

We compare all 8 variants using leave-one-out cross-validation [LOO; Vehtari et al., 2017]. LOO-ELPD estimates each model's predictive accuracy by measuring how well it predicts each observation when that observation is held out, implicitly penalizing model complexity without requiring explicit parameter counting. Table 1 reports the expected log pointwise predictive density (ELPD; higher is better), the

difference relative to the best model (ΔELPD; negative means worse), and the Bayesian stacking weight (the optimal weight for combining models into a predictive ensemble). Harvey curves dominate logistic ($|\Delta\text{ELPD}/\Delta\text{SE}| > 5$), and both the skew-normal likelihood and the joint model usually improve fit. Stacking assigns essentially all weight to the two joint Harvey variants, with the skew-normal variant alone receiving 72% of the stacking weight, and negligible weight to all logistic or independent variants.

Model variant	ELPD	ΔELPD	ΔSE	Weight
Harvey joint (skew)	1475.8	0	—	0.72
Harvey joint (normal)	1465.5	−10.3	7.2	0.28
Harvey indep. (skew)	1437.4	−38.4	7.7	≈0
Logistic joint (skew)	1425.6	−50.2	8.2	0
Harvey indep. (normal)	1420.2	−55.6	9.9	≈0
Logistic joint (normal)	1418.5	−57.3	10.3	≈0
Logistic indep. (normal)	1399.6	−76.2	11.2	≈0
Logistic indep. (skew)	1399.3	−76.5	10.1	0

Table 1: LOO-ELPD model comparison across all 8 variants [Vehtari et al., 2017]. Higher ELPD indicates better predictive fit. ΔELPD is the difference relative to the best model; ΔSE is its standard error. Weight is the Bayesian stacking weight.

B.2 Sensitivity to Saturation Threshold

Table 2 reports the proportion of benchmarks projected to reach saturation by 2030 across all 8 model variants and three threshold definitions (90%, 95%, 99% of score range, from random chance to their estimated upper asymptote L_i). The qualitative finding is robust to all modeling alternatives (≥93% of benchmarks saturated by 2030 at the 95% threshold). Hierarchical (joint) models produce higher saturation estimates, as expected from information sharing across benchmarks. Saturation estimates are lower with the 99% threshold, although, even under this strict definition, 95% of benchmarks are estimated to be saturated by 2030 with the best model, and ≥84% for all model variants.

C Retrodiction Analysis

We validate our forecasting methodology using temporal holdout: training on data before a cutoff date and evaluating predictions on subsequently observed scores.

C.1 Protocol

We set the cutoff date to January 1, 2025. Models are trained on all benchmark data prior to this date, then asked to predict scores observed on models released between January 1, 2025 and January 1, 2026. Benchmarks with fewer pre-cutoff observations than the minimum required in our main dataset are excluded to ensure comparable conditions between the validation and forecasting settings. We compare all model variants.

C.2 Evaluation Metrics

Bayesian Predictive Calibration. We assess whether predicted credible intervals achieve their nominal coverage. For each confidence level $p \in [0.01, 0.99]$, we compute the fraction of test observations falling within the p-credible interval of the posterior predictive distribution. A well-calibrated model shows observed coverage matching expected coverage.

CRPS. The Continuous Ranked Probability Score measures the quality of probabilistic forecasts, penalizing both miscalibration and lack of sharpness. It generalizes the Brier score to continuous predictions. Lower CRPS indicates better predictions.

RMSE. Root Mean Squared Error between the posterior mean prediction and observed scores.

C.3 Retrodiction Results

Supp. Figure 5 showcases model calibration. Full retrodiction metrics (CRPS, RMSE) for all 8 model variants are reported in Table 3.

All eight model variants achieve strong out-of-sample predictive performance (CRPS ≤ 0.065), with good calibration as shown in the calibration curves (Supp. Figure 5). The skew-normal likelihood as well as the joint model consistently improve both CRPS and RMSE, while the Harvey and Logistic models yield similar retro-predictive accuracy.

We select the Harvey hierarchical model with skew-normal likelihood for our main analysis. The hierarchical structure enables information sharing across benchmarks without degrading performance, which may particularly benefit predictions for data-sparse benchmarks. Although the Harvey model is more complex than the logistic, Bayesian inference naturally regularizes through prior specifications, avoiding the overfitting concerns that arise with complex models in classical machine learning. We therefore choose the model that aligns with our theoretical understanding of capability progress and that is confirmed as optimal by formal model comparison (Supp. Table 1).

C.4 Validation: Conformal Prediction Intervals

As a model-free validation of our uncertainty quantification, we apply Conformal Quantile Regression

	Threshold: 90%		Threshold: 95%		Threshold: 99%	
Model variant	Median	80% CI	Median	80% CI	Median	80% CI
Harvey joint (skew)	98.4%	[96.8, 100]	98.4%	[96.8, 100]	95.2%	[90.5, 98.4]
Harvey joint (normal)	100%	[96.8, 100]	98.4%	[95.2, 100]	93.7%	[88.9, 98.4]
Harvey indep. (skew)	95.2%	[93.7, 98.4]	93.7%	[90.5, 96.8]	87.3%	[82.5, 88.9]
Harvey indep. (normal)	95.2%	[93.7, 98.4]	93.7%	[90.5, 96.8]	85.7%	[82.5, 90.5]
Logistic joint (skew)	98.4%	[96.8, 100]	96.8%	[93.7, 98.4]	87.3%	[84.1, 92.1]
Logistic joint (normal)	98.4%	[96.8, 100]	96.8%	[95.2, 100]	90.5%	[85.7, 93.7]
Logistic indep. (skew)	95.2%	[93.7, 98.4]	93.7%	[90.5, 96.8]	84.1%	[81.0, 87.3]
Logistic indep. (normal)	96.8%	[93.7, 98.4]	93.7%	[88.9, 98.4]	84.1%	[79.4, 87.3]

Table 2: Sensitivity of saturation projections across all 8 model variants and three threshold definitions (90%, 95% and 99% of the estimated asymptote).

[CQR; Romano et al., 2019] to the temporal hold-out predictions for all 8 model variants. CQR is a distribution-free method that adjusts Bayesian credible intervals using held-out calibration data: it computes a correction term Q such that the adjusted intervals $[\text{lower} - Q, \text{upper} + Q]$ achieve guaranteed coverage under exchangeability. A value of $Q \approx 0$ indicates that the Bayesian intervals are already well-calibrated and require no correction. Table 3 confirms that corrections are small across all variants (and especially the main one).

D Additional Results

Figure 6 shows trajectories for additional benchmark categories. Commonsense reasoning benchmarks (HellaSwag, PIQA, WinoGrande) saturated earliest, consistent with these tasks representing lower cognitive complexity. Language understanding and multimodal benchmarks show intermediate trajectories, with saturation projected by 2028-2029.

E Human Performance Baselines

Table 4 summarizes human performance baselines used in our analysis. We categorize human performance into four groups based on expertise level:

- **Average Human**: Crowdworkers (e.g., MTurk) or non-specialized participants
- **Skilled Generalist**: Individuals with advanced education but not in the target domain (e.g., PhD students in unrelated fields, skilled professionals)
- **Domain Expert**: PhD-level specialists in the relevant domain or expert professionals
- **Top Performer**: Elite performers (e.g., top 5% test takers or best result)

Committee scores represent an aggregation of majority votes or average team scores across human participants. The *High School Qualifier* and *High School Top Performer* categories specifically represent a cohort of students in a specialized training program.

Interpretation notes. Human baselines vary substantially depending on expertise level and evaluation conditions. For instance, on GPQA Diamond, the gap between skilled generalists (22%) and domain experts (81%) spans nearly 60 percentage points, illustrating how benchmark difficulty depends critically on domain knowledge. Similarly, on mathematical benchmarks, the gap between a CS PhD student (40% on MATH Level 5) and an elite mathematician (90%) reflects the specialized nature of competition-level mathematics.

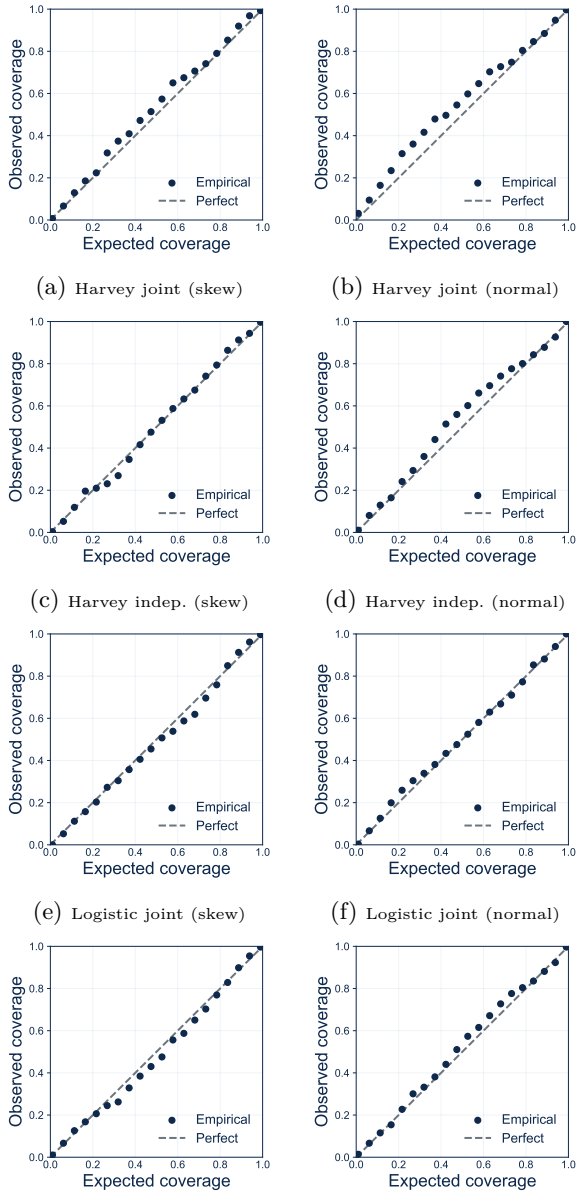
These baselines provide context for interpreting AI progress: when frontier models exceed domain expert performance, as they now do on several benchmarks (GPQA Diamond, MMLU, HellaSwag), this represents a meaningful capability threshold. However, we caution that benchmark performance may not directly translate to real-world task completion, and that human baselines themselves have limitations (small sample sizes, varying incentive structures, potential ceiling effects).

Model variant	Bayes. cov.	CQR cov.	Q	CRPS	RMSE
Harvey joint (skew)	83.2%	83.2%	−0.000	**0.054**	**0.104**
Harvey joint (normal)	83.9%	88.1%	+0.012	0.058	0.108
Harvey indep. (skew)	79.7%	69.9%	−0.010	0.055	0.107
Harvey indep. (normal)	81.1%	83.2%	+0.005	0.064	0.118
Logistic joint (skew)	77.6%	79.7%	+0.002	**0.054**	**0.103**
Logistic joint (normal)	83.2%	90.9%	+0.013	0.058	0.108
Logistic indep. (skew)	76.2%	71.3%	−0.005	**0.054**	**0.104**
Logistic indep. (normal)	81.8%	83.2%	+0.008	0.063	0.114

Table 3: Retrodiction results and CQR calibration across all 8 model variants. CQR is evaluated at the 80% nominal level ($n_{\text{cal}} = 143$, $n_{\text{test}} = 143$); the conformal adjustment Q quantifies how much the Bayesian intervals (Bayes cov.) need correction ($Q \approx 0$ indicates well-calibrated intervals). CRPS and RMSE are out-of-sample retrodiction metrics (lower is better). The main model (Harvey joint, skew-normal) achieves 83.2% Bayesian coverage with essentially no conformal correction ($Q \approx 0$). Bold CRPS/RMSE values are statistically indistinguishable from the best.

Benchmark	Human Group	Score	Source
Domain-Specific Knowledge			
GPQA Diamond	Domain Expert	81.2%	Rein et al. [2023]
GPQA Diamond	Domain Expert	69.7%	OpenAI [2024]
GPQA Diamond	Skilled Generalist	21.9%	Rein et al. [2023]
PRBench Legal	Committee of Domain Experts	79.6%	Akyürek et al. [2025]
PRBench Finance	Committee of Domain Experts	79.6%	Akyürek et al. [2025]
GSM8K	Skilled Generalist	96.8%	Li et al. [2024b]
ScienceQA	Average Human	88.4%	Lu et al. [2022]
General Reasoning			
SimpleBench	Average Human	83.7%	Philip and Hemang [2024]
Mathematical Reasoning			
MATH Level 5	Domain Expert	90%	Hendrycks et al. [2021]
MATH Level 5	Skilled Generalist	40%	Hendrycks et al. [2021]
FrontierMath	Committee of Domain Experts	35%	Epoch AI [2025c]
OTIS Mock AIME	High School Qualifier	53%	Chen [2025]
OTIS Mock AIME	High School Top Performer	93%	Chen [2025]
AGI Progress			
ARC-AGI-1	Average Human	77%	ARC Prize [2019]
ARC-AGI-1	Committee of Average Humans	98%	ARC Prize [2019]
ARC-AGI-1	Skilled Generalist	98%	ARC Prize [2019]
ARC-AGI-2	Average Human	60%	ARC Prize [2025a]
ARC-AGI-2	Committee of Average Humans	100%	ARC Prize [2025a]
Agentic Computer Use			
OS World	Skilled Generalist	72.4%	Xie et al. [2024]
Biology & Chemistry (Dual-Use)			
GPQA Diamond Biology	Skilled Generalist	22.0%	Dev et al. [2025]
GPQA Diamond Biology	Domain Expert	83.1%	Dev et al. [2025]
GPQA Diamond Chemistry	Skilled Generalist	22.0%	Dev et al. [2025]
GPQA Diamond Chemistry	Domain Expert	83.1%	Dev et al. [2025]
WMDP Biology	Domain Expert	60%	Dev et al. [2025]
WMDP Chemistry	Domain Expert	43%	Dev et al. [2025]
LAB-Bench Cloning	Domain Expert	60%	Dev et al. [2025]
LAB-Bench LitQA2	Domain Expert	70%	Dev et al. [2025]
LAB-Bench Protocol	Domain Expert	79%	Dev et al. [2025]
LAB-Bench SeqQA	Domain Expert	78%	Dev et al. [2025]
BioLP-bench	Domain Expert	38%	Dev et al. [2025]
Commonsense Reasoning			
HellaSwag	Committee of Average Humans	95.6%	Zellers et al. [2019]
WinoGrande	Committee of Average Humans	94.0%	Sakaguchi et al. [2020]
PIQA	Committee of Skilled Generalists	94.9%	Bisk et al. [2020]
OpenBookQA	Average Human	92%	Mihaylov et al. [2018]
Multimodal understanding			
GeoBench	Top Performer	90%	ccmdi [2026]
VISTA	Skilled Generalist	55.4%	Scale AI [2025]
VPCT	Average Human	100%	Brower [2025]

Table 4: **Human performance baselines by benchmark.** Scores represent accuracy unless otherwise noted. Committee scores use majority votes or average team scores across human participants. Details regarding each human baseline are available in the supplementary material.



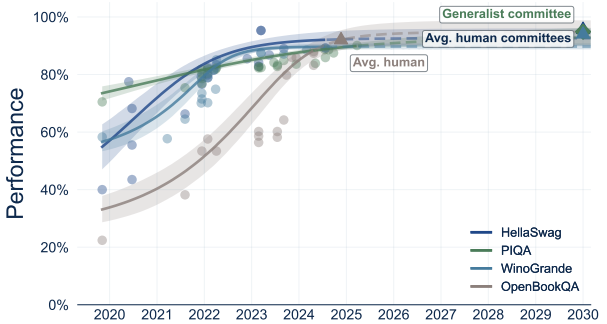
(a) Harvey joint (skew) (b) Harvey joint (normal)

(c) Harvey indep. (skew) (d) Harvey indep. (normal)

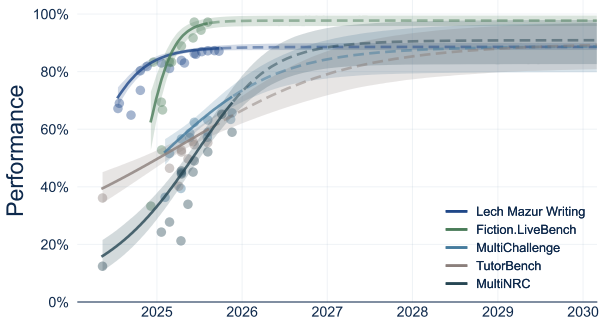
(e) Logistic joint (skew) (f) Logistic joint (normal)

(g) Logistic indep. (skew) (h) Logistic indep. (normal)

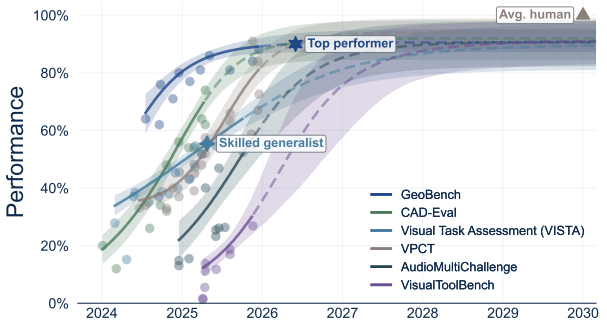
Figure 5: Calibration curves for all 8 model variants. Well-calibrated models produce points close to the diagonal. All variants achieve reasonable calibration, with no systematic difference between modeling choices.



(a) Commonsense reasoning



(b) Language understanding



(c) Multimodal understanding

Figure 6: **Additional benchmark trajectories.** Additional capability categories showing consistent saturation patterns. Commonsense reasoning benchmarks (HellaSwag, PIQA, WinoGrande) are already near saturation; language and multimodal understanding show trajectories consistent with other categories.