Position: The Perspective of Superhuman AI on Most Cognitive Tasks by 2030 Warrants Immediate Preemptive Action

Anonymous Authors[1]

Abstract

In this position paper, we argue that without coordinated intervention, AI systems will by default exceed human expert performance on most cognitive tasks by 2030. Using a joint hierarchical Bayesian model, we forecast frontier performance trajectories across 60 benchmarks spanning reasoning, mathematics, coding, scientific knowledge, and agentic capabilities. We find that nearly all current benchmarks (98% in our set) are on track to saturate within four years. Security-critical benchmarks (cybersecurity, autonomous AI R&D, biology, chemistry) show even faster trajectories, saturating before 2028. While these findings may not constitute definitive proof of imminent AGI on their own, they 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.

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 exceed human expert performance on most measurable cognitive tasks by 2030. This projection, which includes security-critical capabilities, leaves a limited runway before irreversible impacts, thereby warranting pressing preemptive 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 limited runway for developing robust alignment or control techniques, given the current 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 60 benchmarks showing convergent saturation trajectories based on a new modeling framework; (2) We analyze AI progress specifically in security-critical capability domains; (3) We propose concrete actions for researchers and policymakers.

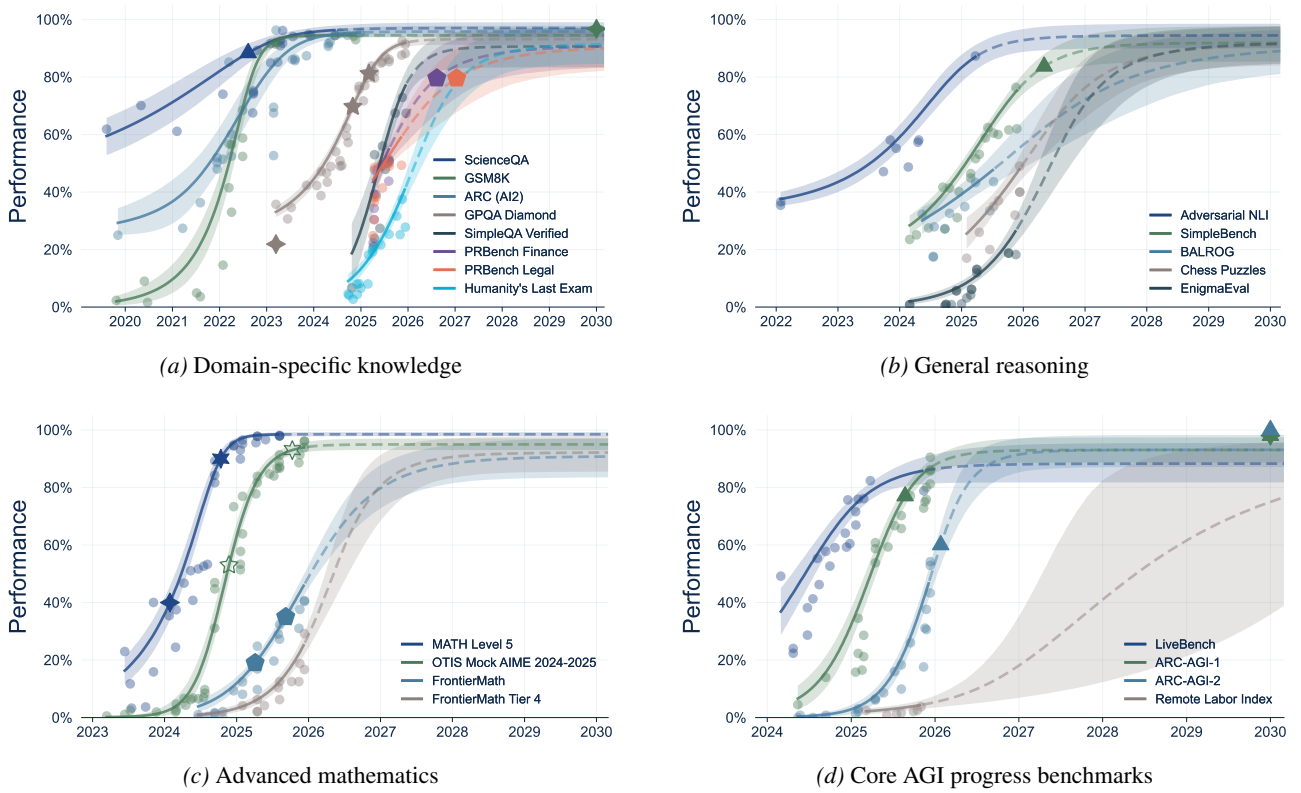
2. Evidence: Benchmark Trajectories

2.1. Data and Methodology

We analyze performance trajectories on 60 benchmarks spanning diverse cognitive capabilities: commonsense,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.



(a) Domain-specific knowledge

(b) General reasoning

(c) Advanced mathematics

(d) Core 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 D). All categories show rapid progress towards saturation, with several benchmarks already exceeding human expert baselines.

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 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.

(2) Hierarchical Bayesian modeling. Rather than fitting each benchmark independently, we jointly model all 60 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.

(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 validate our methodology through temporal holdout: training on data before January 2025 and evaluating pre-

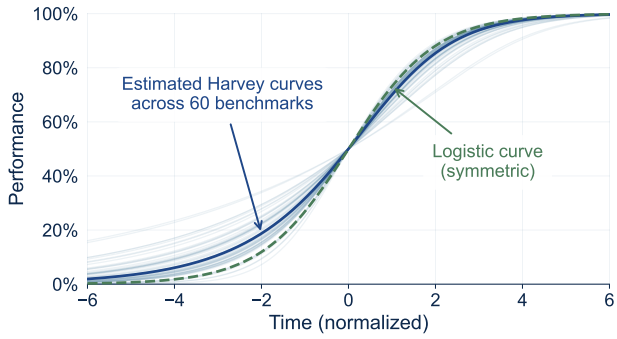


Figure 2. **Harvey curves capture asymmetric progress trajectories.** Fitted Harvey curves (blue) for 60 benchmarks show gradual acceleration followed by rapid saturation, compared to the symmetric logistic curve (green dashed). All curves are standardized to reach 50% performance at time 0 and have a growth rate of 1, in order to compare shapes only. The thick blue curve is the median Harvey trajectory.

dictions on subsequently observed scores. Hierarchical and independent fits as well as Harvey and logistic models achieve similar predictive accuracy (CRPS, RMSE), but the more general Harvey curves are preferred in this article as they better align with our theoretical understanding of capability progress. Unlike classical machine learning, Bayesian inference regularizes through prior specifications, avoiding overfitting despite the model's greater complexity. See Appendix A for model specification and Appendix B for validation details.

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).

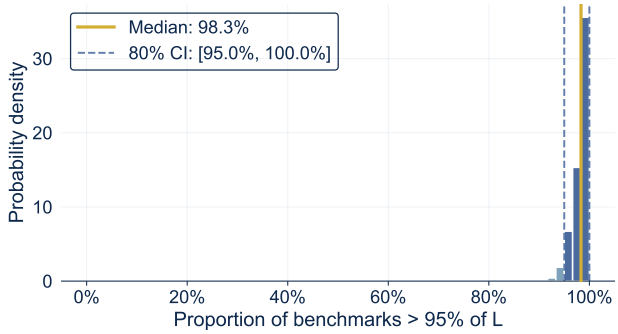


Figure 3. **Nearly all benchmarks projected to saturate by 2030.** Posterior distribution of the proportion of benchmarks reaching 95% of their maximum score range by 2030. The median is 98.3%, with 80% credible interval [95.0%, 100%], indicating that saturation of most benchmarks by 2030 is the default scenario, barring exogenous intervention.

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 D.

- **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 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 & Hemang, 2024; Paglieri et al., 2025; Epoch AI, 2025b; 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, 2025f;e). 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 top performer (Fields Medal-level mathematician) achieves 90% (Hendrycks et al., 2021); frontier models now approach the latter. On FrontierMath, teams of mathematicians collectively solve only 19–35% of problems in four and a half hours, with internet access (Epoch AI, 2025e). Our analysis projects rapid progress and saturation in 2028.

- **Core AGI progress** (LiveBench, ARC-AGI v1 and v2, soon v3, Remote Labor Index) (White et al., 2025; Chollet, 2019; ARC Prize, 2019; 2025a;b; Mazeika 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 real-world tasks from online freelance platforms. LiveBench and ARC-AGI-2 should saturate soon; 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) (aider, 2024; Kwa et al., 2025; Ihle, 2025; Epoch AI, 2024b; Jimenez et al., 2024; Deng et al., 2025; Shetty et al., 2025). Benchmarks measuring autonomous software engineering and ML research capabilities. Saturation projected by 2028, 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 D); frontier models already match or exceed these baselines on several subtasks. Saturation projected by 2027-2028.

Three additional benchmark categories are presented in Appendix C: **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, 2025a; 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

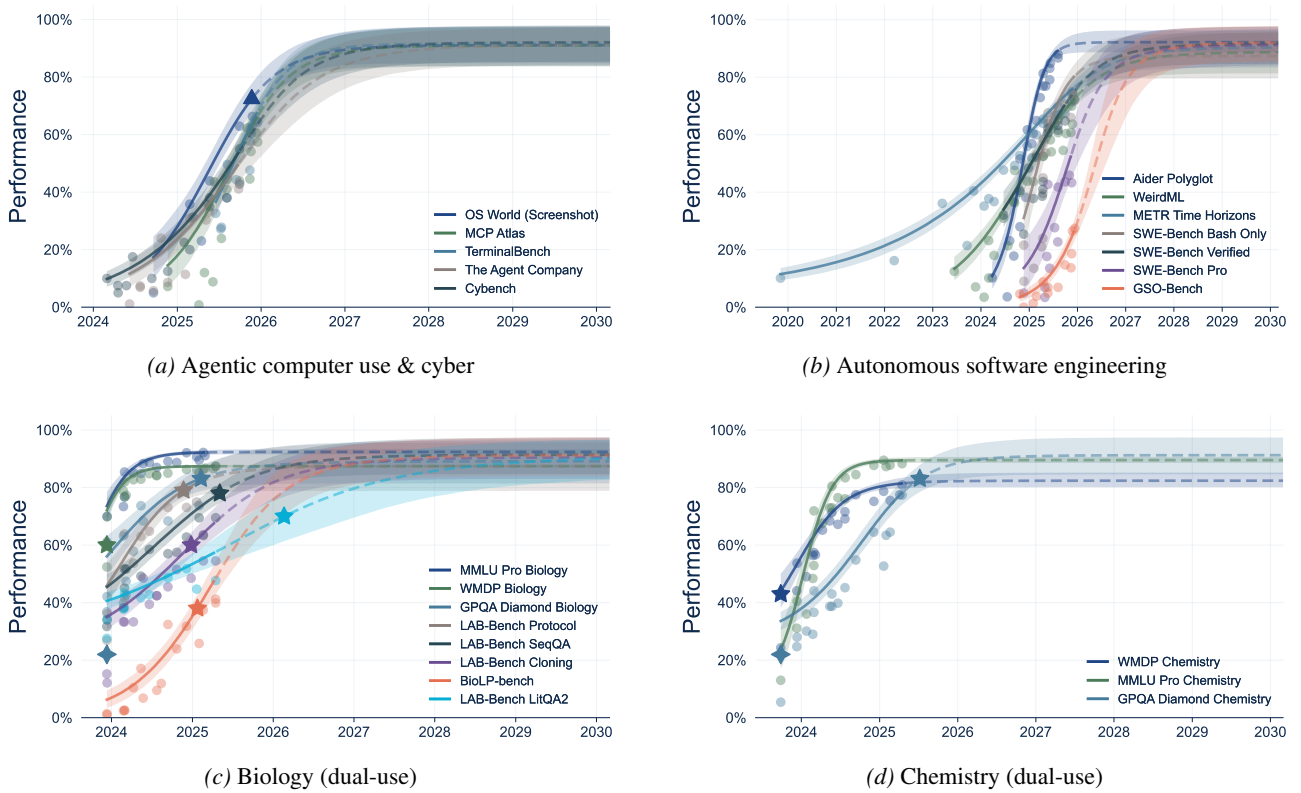
Our projections should be interpreted against a recent pattern of systematic underestimation of AI progress (Kučinskas et al., 2025). This contrasts with the earlier history of AI, with periods of excessive optimism about symbolic AI and expert systems, followed by "AI winters" in the 1970s and late 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, 2025c), and research priorities (Erdil & 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.



(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.

3.2. 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 & 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.

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 narrow optimization by AI companies incentivized to showcase impressive results with each new model release (Balloccu et al., 2024; Robison, 2025).

Response: We acknowledge these concerns but note several mitigating factors: (1) The saturation trends presented in this article are consistent across 60 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; 2025e). (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.

4.2. 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.

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) the breadth of saturation across diverse tasks is difficult to explain by narrow optimization alone.

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

4.3. 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.

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 (test-time compute, reasoning models) continue to unlock progress, 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).

4.4. Heterogeneous Progress Blocks AGI

Objection: AI progress is "jagged", superhuman on some tasks yet subhuman on others. This heterogeneity may persist, preventing AGI-like systems.

Response: We acknowledge heterogeneity but note that (1) gaps are progressively closing across diverse capabilities, (2) as explained in Section 4.2, jaggedness does not preclude transformative impact, including critical and irreversible ones, 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.

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;
- Demonstration that benchmark performance system-

atically 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.

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 leads to superhuman performance on most measurable cognitive tasks by 2030. These trajectories are not inevitable, as they result from choices that can still be influenced through coordinated action, but the window is closing rapidly. Given the state of our understanding of and control over these systems, the cost of inaction could be severe.

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 [ANONYMOUS URL].

References

Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T., and Zinkov, R. 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.

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

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.

Akyürek, A. F., Gosai, A., Zhang, C. B. C., Gupta, V., Jeong, J., Gunjal, A., Rabbani, T., Mazzone, M., Randolph, D., Meymand, M. M., Chattha, G., Rodriguez, P., Mares, D., Singh, P., Liu, M., Chawla, S., Cline, P., Ogaz, L., Hernandez, E., Wang, Z., Bhatter, P., Ayestaran, M., Liu, B., and He, Y. 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.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete Problems in AI Safety, July 2016. URL http://arxiv.org/abs/1606.06565. arXiv:1606.06565 [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/.

Balloccu, S., Schmidtová, P., Lango, M., and Dušek, O. 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].

Bandi, C., Hertzberg, B., Boo, G., Polakam, T., Da, J., Hassaan, S., Sharma, M., Park, A., Hernandez, E., Rambado, D., Salazar, I., Rane, C., Levin, B., Kenstler, B., and Liu, B. MCP-Atlas: A Large-Scale Benchmark for Tool-Use Competency with Real MCP Servers, December 2025.

Barrett, S., Murray, M., Quarks, O., Smith, M., Kryś, J., Campos, S., Boria, A. T., Touzet, C., Hayrapet, S., Heiding, F., Nevo, O., Swanda, A., Aguirre, J., Gershovich, A. B., Clay, E., Fetterman, R., Fritz, M., Juarez, M., Mavroudis, V., and Papadatos, H. Toward Quantitative Modeling of Cybersecurity Risks Due to AI Misuse, December 2025. URL http://arxiv.org/abs/2512.08864. arXiv:2512.08864 [cs].

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., Newman, J., Ng, K. Y., Okolo, C. T., Raji, D., Sastry, G., Seger, E., Skeadas, T., South, T., Strubell, E., Tramèr, F., Velasco, L., Wheeler, N., Acemoglu, D., Adekanmbi, O., Dalrymple, D., Dietterich, T. G., Felten, E. W., Fung, P., Gourinchas, P.-O., Heintz, F., Hinton, G., Jennings, N., Krause, A., Leavy, S., Liang, P., Ludermir, T., Marda, V., Margetts, H., McDermid, J., Munga, J., Narayanan, A., Nelson, A., Neppel, C., Oh, A., Ramchurn, G., Russell, S., Schaake, M., Schölkopf, B., Song, D., Soto, A., Tiedrich, L., Varoquaux, G., Yao, A., Zhang, Y.-Q., Albalawi, F., Alserkal, M., Ajala, O., Avrin, G., Busch, C., Carvalho, A. C. P. d. L. F. d., Fox, B., Gill, A. S., Hatip, A. H., Heikkilä, J., Jolly, G., Katzir, Z., Kitano, H., Krüger, A., Johnson, C., Khan, S. M., Lee, K. M., Ligot, D. V., Molchanovskyi, O., Monti, A., Mwamanzi, N., Nemer, M., Oliver, N., Portillo, J. R. L., Ravindran, B., Rivera, R. P., Riza, H., Rugege, C., Seoighe, C., Sheehan, J., Sheikh, H., Wong, D., and Zeng, Y. International AI Safety Report. Technical Report arXiv:2501.17805, arXiv, January 2025. URL http://arxiv.org/abs/2501.17805. arXiv:2501.17805 [cs].

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

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

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

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, September 2023. URL http://arxiv.org/abs/2307.15217. arXiv:2307.15217 [cs].

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

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

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

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. 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].

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training Verifiers to Solve Math Word Problems, November 2021.

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

Dalrymple, D. d., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., Abate, A., Halpern, J., Barrett, C., Zhao, D., Zhi-Xuan, T., Wing, J., and Tenenbaum, J. 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].

Deng, X., Da, J., Pan, E., He, Y. Y., Ide, C., Garg, K., Lauffer, N., Park, A., Pasari, N., Rane, C., Sampath, K., Krishnan, M., Kundurthy, S., Hendryx, S., Wang, Z., Bharadwaj, V., Holm, J., Aluri, R., Zhang, C. B. C., Jacobson, N., Liu, B., and Kenstler, B. 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].

Dev, S., Teague, C., Ellison, G., Brady, K., Lee, Y.-C. J., Gebauer, S. L., Bradley, H. A., Maciorowski, D., Persaud, B., Despanie, J., Del Castello, B., Worland, A., Miller, M., Salas, A., Nguyen, D., Liu, J., Johnson, J., Sloan, A., Stonehouse, W., Merrill, T., Goode, T., McKelvey, G., and Guest, E. 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.

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

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

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

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

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

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

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

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

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

Fabbri, A. R., Mares, D., Flores, J., Mankikar, M., Hernandez, E., Lee, D., Liu, B., and Xing, C. MultiNRC: A Challenging and Native Multilingual Reasoning Evaluation Benchmark for LLMs, July 2025. URL http://arxiv.org/abs/2507.17476. arXiv:2507.17476 [cs].

Gosai, A., Vuong, T., Tyagi, U., Li, S., You, W., Bavare, M., Uçar, A., Fang, Z., Jang, B., Liu, B., and He, Y. 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].

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O. When Will AI Exceed Human Performance? Evidence from AI Experts, May 2018. URL http://arxiv.org/abs/1705.08807. arXiv:1705.08807 [cs].

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., Brauner, J., and Korzekwa, R. C. 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].

Guo, X., Tyagi, U., Gosai, A., Vergara, P., Park, J., Montoya, E. G. H., Zhang, C. B. C., Hu, B., He, Y., Liu, B., and Srinivasa, R. S. 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].

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

Harvey, A. C. 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.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring

Mathematical Problem Solving With the MATH Dataset, November 2021. URL http://arxiv.org/abs/2103.03874. arXiv:2103.03874 [cs].

Ho, A., Denain, J.-S., Atanasov, D., Albanie, S., and Shah, R. A Rosetta Stone for AI Benchmarks, November 2025. URL http://arxiv.org/abs/2512.00193. arXiv:2512.00193 [cs].

Ihle, H. T. WeirdML, 2025. URL https://htihle.github.io/weirdml.html.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. R. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *Proc. of ICLR*. OpenReview.net, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs].

Kipnis, A., Voudouris, K., Buschoff, L. M. S., and Schulz, E. 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].

Kučinskas, S., Rosenberg, J., de Castro, R. C., Jacobs, Z., Canedy, J., Tetlock, P. E., and Karger, E. 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.

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., Rein, D., Sato, L. J. K., Wijk, H., Ziegler, D. M., Barnes, E., and Chan, L. Measuring AI Ability to Complete Long Tasks, March 2025. URL http://arxiv.org/abs/2503.14499. arXiv:2503.14499 [cs].

Laurent, J. M., Janizek, J. D., Ruzo, M., Hinks, M. M., Hammerling, M. J., Narayanan, S., Ponnapati, M., White, A. D., and Rodriques, S. G. LAB-Bench: Measuring Capabilities of Language Models for Biology Research, July 2024. URL http://arxiv.org/abs/2407.10362. arXiv:2407.10362 [cs].

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J. D., Dombrowski, A.-K., Goel, S., Phan, L., Mukobi, G., Helm-Burger, N., Lababidi, R., Justen, L., Liu, A. B., Chen, M., Barrass, I., Zhang, O., Zhu, X., Tamirisa, R., Bharathi, B., Khoja, A., Zhao, Z., Herbert-Voss, A., Breuer, C. B., Marks, S., Patel, O., Zou, A., Mazeika, M., Wang, Z., Oswal, P., Lin, W., Hunt, A. A., Tienken-Harder, J., Shih, K. Y., Talley, K., Guan, J., Kaplan, R., Steneker, I., Campbell, D., Jokubaitis, B., Levinson, A., Wang, J., Qian, W., Karmakar, K. K., Basart, S., Fitz, S., Levine, M., Kumaraguru, P., Tupakula, U., Varadharajan, V., Wang, R., Shoshitaishvili, Y., Ba, J., Esvelt, K. M., Wang, A., and Hendrycks, D. The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning, May 2024a. URL http://arxiv.org/abs/2403.03218. arXiv:2403.03218 [cs].

Li, Q., Cui, L., Zhao, X., Kong, L., and Bi, W. 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].

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. 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].

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

Mazeika, M., Gatti, A., Menghini, C., Sehwag, U. M., Singhal, S., Orlovskiy, Y., Basart, S., Sharma, M., Peskoff, D., Lau, E., Lim, J., Carroll, L., Blair, A., Sivakumar, V., Basu, S., Kenstler, B., Ma, Y., Michael, J., Li, X., Ingebretsen, O., Mehta, A., Mottola, J., Teichmann, J., Yu, K., Shaik, Z., Khoja, A., Ren, R., Hausenloy, J., Phan, L., Htet, Y., Aich, A., Rabbani, T., Shah, V., Novykov, A., Binder, F., Chugunov, K., Ramirez, L., Geralnik, M., Mesura, H., Lee, D., Cardona, E.-Y. H., Diamond, A., Yue, S., Wang, A., Liu, B., Hernandez, E., and Hendrycks, D. Remote Labor Index: Measuring AI Automation of Remote Work, October 2025. URL http://arxiv.org/abs/2510.26787. arXiv:2510.26787 [cs].

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

Merrill, M. A., Shaw, A. G., Carlini, N., Li, B., Raj, H., Bercovich, I., Shi, L., Shin, J. Y., Walshe, T., Buchanan, E. K., Shen, J., Ye, G., Lin, H., Poulos, J., Wang, M., Nezhurina, M., Jitsev, J., Lu, D., Mastromichalakis,

O. M., Xu, Z., Chen, Z., Liu, Y., Zhang, R., Chen, L. L., Kashyap, A., Uslu, J.-L., Li, J., Wu, J., Yan, M., Bian, S., Sharma, V., Sun, K., Dillmann, S., Anand, A., Lanpouthakoun, A., Koopah, B., Hu, C., Guha, E., Dreiman, G. H. S., Zhu, J., Krauth, K., Zhong, L., Muennighoff, N., Amanfu, R., Tan, S., Pimpalgaonkar, S., Aggarwal, T., Lin, X., Lan, X., Zhao, X., Liang, Y., Wang, Y., Wang, Z., Zhou, C., Heineman, D., Liu, H., Trivedi, H., Yang, J., Lin, J., Shetty, M., Yang, M., Omi, N., Raoof, N., Li, S., Zhuo, T. Y., Lin, W., Dai, Y., Wang, Y., Chai, W., Zhou, S., Wahdany, D., She, Z., Hu, J., Dong, Z., Zhu, Y., Cui, S., Saiyed, A., Kolbeinsson, A., Hu, J., Rytting, C. M., Marten, R., Wang, Y., Dimakis, A., Konwinski, A., and Schmidt, L. 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].

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. 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].

Murray, M., Barrett, S., Papadatos, H., Quarks, O., Smith, M., Boria, A. T., Touzet, C., and Campos, S. A Methodology for Quantitative AI Risk Modeling, December 2025. URL http://arxiv.org/abs/2512.08844. arXiv:2512.08844 [cs].

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

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial NLI: A New Benchmark for Natural Language Understanding. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proc. of ACL*, pp. 4885–4901, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

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

Paglieri, D., Cupiał, B., Coward, S., Piterbarg, U., Wolczyk, M., Khan, A., Pignatelli, E., Kuciński, L., Pinto, L., Fergus, R., Foerster, J. N., Parker-Holder, J., and Rocktäschel, T. BALROG: Benchmarking Agentic LLM and VLM Reasoning On Games, April 2025. URL http://arxiv.org/abs/2411.13543. arXiv:2411.13543 [cs].

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., Dodonov, D., Nguyen, T., Lee, J., Anderson, D., Doroshenko, M., Stokes, A. C., Mahmood, M., Pokutnyi, O., Iskra, O., Wang, J. P., Levin, J.-C., Kazakov, M., Feng, F., Feng, S. Y., Zhao, H., Yu, M., Gangal, V., Zou, C., Wang, Z., Popov, S., Gerbicz, R., Galgon, G., Schmitt, J., Yeadon, W., Lee, Y., Sauers, S., Sanchez, A., Giska, F., Roth, M., Riis, S., Utpala, S., Burns, N., Goshu, G. M., Naiya, M. M., Agu, C., Giboney, Z., Cheatom, A., Fournier-Facio, F., Crowson, S.-J., Finke, L., Cheng, Z., Zampese, J., Hoerr, R. G., Nandor, M., Park, H., Gehrunger, T., Cai, J., McCarty, B., Garretson, A. C., Taylor, E., Sileo, D., Ren, Q., Qazi, U., Li, L., Nam, J., Wydallis, J. B., Arkhipov, P., Shi, J. W. L., Bacho, A., Willcocks, C. G., Cao, H., Motwani, S., Santos, E. d. O., Veith, J., Vendrow, E., Cojoc, D., Zenitani, K., Robinson, J., Tang, L., Li, Y., Vendrow, J., Fraga, N. W., Kuchkin, V., Maksimov, A. P., Marion, P., Efremov, D., Lynch, J., Liang, K., Mikov, A., Gritsevskiy, A., Guillod, J., Demir, G., Martinez, D., Pageler, B., Zhou, K., Soori, S., Press, O., Tang, H., Rissone, P., Green, S. R., Brüssel, L., Twayana, M., Dieuleveut, A., Imperial, J. M., Prabhu, A., Yang, J., Crispino, N., Rao, A., Zvonkine, D., Loiseau, G., Kalinin, M., Lukas, M., Manolescu, C., Stambaugh, N., Mishra, S., Hogg, T., Bosio, C., Coppola, B. P., Salazar, J., Jin, J., Sayous, R., Ivanov, S., Schwaller, P., Senthilkuma, S., Bran, A. M., Algaba, A., Houte, K. V. d., Sypt, L. V. D., Verbeken, B., Noever, D., Kopylov, A., Myklebust, B., Li, B., Schut, L., Zheltonozhskii, E., Yuan, Q., Lim, D., Stanley, R., Yang, T., Maar, J., Wykowski, J., Oller, M., Sahu, A., Ardito, C. G., Hu, Y., Kamdoum, A. G. K., Jin, A., Vilchis, T. G., Zu, Y., Lackner, M., Koppel, J., Sun, G., Antonenko, D. S., Chern, S., Zhao, B., Arsene, P., Cavanagh, J. M., Li, D., Shen, J., Crisostomi, D., Zhang, W., Dehghan, A., Ivanov, S., Perrella, D., Kaparov, N., Zang, A., Sucholutsky, I., Kharlamova, A., Orel, D., Poritski, V., Ben-David, S., Berger, Z., Whitfill, P., Foster, M., Munro, D., Ho, L., Sivarajan, S., Hava, D. B., Kuchkin, A., Holmes, D., Rodriguez-Romero, A., Sommerhage, F., Zhang, A., Moat, R., Schneider, K., Kazibwe, Z., Clarke, D., Kim, D. H., Dias, F. M., Fish, S., Elser, V., Kreiman, T., Vilchis, V. E. G., Klose, I., Anantheswaran, U., Zweiger, A., Rawal, K., Li, J., Nguyen, J., Daans, N., Heidinger, H., Radionov, M., Rozhoň, V., Ginis, V., Stump, C., Cohen, N., Poświata, R., Tkadlec, J., Goldfarb, A., Wang, C., Padlewski, P., Barzowski, S., Montgomery, K., Stendall, R., Tucker-Foltz, J., Stade, J., Rogers, T. R., Goertzen, T., Grabb, D., Shukla, A., Givré, A., Ambay, J. A., Sen, A., Aziz, M. F., Inlow, M. H., He, H., Zhang, L., Kaddar, Y., Ängquist, I., Chen, Y., Wang, H. K., Ramakrish-

nan, K., Thornley, E., Terpin, A., Schoelkopf, H., Zheng, E., Carmi, A., Brown, E. D. L., Zhu, K., Bartolo, M., Wheeler, R., Stehberger, M., Bradshaw, P., Heimonen, J. P., Sridhar, K., Akov, I., Sandlin, J., Makarychev, Y., Tam, J., Hoang, H., Cunningham, D. M., Goryachev, V., Patramanis, D., Krause, M., Redenti, A., Aldous, D., Lai, J., Coleman, S., Xu, J., Lee, S., Magoulas, I., Zhao, S., Tang, N., Cohen, M. K., Paradise, O., Kirchner, J. H., Ovchynnikov, M., Matos, J. O., Shenoy, A., Wang, M., Nie, Y., Sztyber-Betley, A., Faraboschi, P., Riblet, R., Crozier, J., Halasyamani, S., Verma, S., Joshi, P., Meril, E., Ma, Z., Andréoletti, J., Singhal, R., Platnick, J., Nevirkovets, V., Basler, L., Ivanov, A., Khoury, S., Gustafsson, N., Piccardo, M., Mostaghimi, H., Chen, Q., Singh, V., Khánh, T. Q., Rosu, P., Szlyk, H., Brown, Z., Narayan, H., Menezes, A., Roberts, J., Alley, W., Sun, K., Patel, A., Lamparth, M., Reuel, A., Xin, L., Xu, H., Loader, J., Martin, F., Wang, Z., Achilleos, A., Preu, T., Korbak, T., Bosio, I., Kazemi, F., Chen, Z., Bálint, B., Lo, E. J. Y., Wang, J., Nunes, M. I. S., Milbauer, J., Bari, M. S., Wang, Z., Ansarinejad, B., Sun, Y., Durand, S., Elgnainy, H., Douville, G., Tordera, D., Balabanian, G., Wolff, H., Kvistad, L., Milliron, H., Sakor, A., Eron, M., O, A. F. D., Shah, S., Zhou, X., Kamalov, F., Abdoli, S., Santens, T., Barkan, S., Tee, A., Zhang, R., Tomasiello, A., Luca, G. B. D., Looi, S.-Z., Le, V.-K., Kolt, N., Pan, J., Rodman, E., Drori, J., Fossum, C. J., Muennighoff, N., Jagota, M., Pradeep, R., Fan, H., Eicher, J., Chen, M., Thaman, K., Merrill, W., Firsching, M., Harris, C., Ciobâcă, S., Gross, J., Pandey, R., Gusev, I., Jones, A., Agnihotri, S., Zhelnov, P., Mofayezi, M., Piperski, A., Zhang, D. K., Dobarskyi, K., Leventov, R., Soroko, I., Duersch, J., Taamazyan, V., Ho, A., Ma, W., Held, W., Xian, R., Zebaze, A. R., Mohamed, M., Leser, J. N., Yuan, M. X., Yacar, L., Lengler, J., Olszewska, K., Fratta, C. D., Oliveira, E., Jackson, J. W., Zou, A., Chidambaram, M., Manik, T., Haffenden, H., Stander, D., Dasouqi, A., Shen, A., Golshani, B., Stap, D., Kretov, E., Uzhou, M., Zhidkovskaya, A. B., Winter, N., Rodriguez, M. O., Lauff, R., Wehr, D., Tang, C., Hossain, Z., Phillips, S., Samuele, F., Ekström, F., Hammon, A., Patel, O., Farhidi, F., Medley, G., Mohammadzadeh, F., Peñaflor, M., Kassahun, H., Friedrich, A., Perez, R. H., Pyda, D., Sakal, T., Dhamane, O., Mirabadi, A. K., Hallman, E., Okutsu, K., Battaglia, M., Maghsoudimehrabani, M., Amit, A., Hulbert, D., Pereira, R., Weber, S., Handoko, Peristyy, A., Malina, S., Mehkary, M., Aly, R., Reidegeld, F., Dick, A.-K., Friday, C., Singh, M., Shapourian, H., Kim, W., Costa, M., Gurdogan, H., Kumar, H., Ceconello, C., Zhuang, C., Park, H., Carroll, M., Tawfeek, A. R., Steinerberger, S., Aggarwal, D., Kirchhof, M., Dai, L., Kim, E., Ferret, J., Shah, J., Wang, Y., Yan, M., Burdzy, K., Zhang, L., Franca, A., Pham, D. T., Loh, K. Y., Robinson, J., Jackson, A., Giordano, P., Petersen, P., Cosma, A., Colino, J., White, C., Votava, J., Vinnikov, V., Delaney, E., Spelda, P., Stritecky, V., Shahid, S. M., Mourrat, J.-C., Vetoshkin, L., Sponselee, K., Bacho, R., Yong, Z.-X., Rosa, F. d. l., Cho, N., Li, X., Malod, G., Weller, O., Albani, G., Lang, L., Laurendeau, J., Kazakov, D., Adesanya, F., Portier, J., Hollom, L., Souza, V., Zhou, Y. A., Degorre, J., Yalın, Y., Obikoya, G. D., Rai, Bigi, F., Boscá, M. C., Shumar, O., Bacho, K., Recchia, G., Popescu, M., Shulga, N., Tanwie, N. M., Lux, T. C. H., Rank, B., Ni, C., Brooks, M., Yakimchyk, A., Huanxu, Liu, Cavalleri, S., Häggström, O., Verkama, E., Newbould, J., Gundlach, H., Brito-Santana, L., Amaro, B., Vajipey, V., Grover, R., Wang, T., Kratish, Y., Li, W.-D., Gopi, S., Caciolai, A., Witt, C. S. d., Hernández-Cámara, P., Rodolà, E., Robins, J., Williamson, D., Cheng, V., Raynor, B., Qi, H., Segev, B., Fan, J., Martinson, S., Wang, E. Y., Hausknecht, K., Brenner, M. P., Mao, M., Demian, C., Kassani, P., Zhang, X., Avagian, D., Scipio, E. J., Ragoler, A., Tan, J., Sims, B., Plecnik, R., Kirtland, A., Bodur, O. F., Shinde, D. P., Labrador, Y. C. L., Adoul, Z., Zekry, M., Karakoc, A., Santos, T. C. B., Shamseldeen, S., Karim, L., Liakhovitskaia, A., Resman, N., Farina, N., Gonzalez, J. C., Maayan, G., Anderson, E., Pena, R. D. O., Kelley, E., Mariji, H., Pouriamanesh, R., Wu, W., Finocchio, R., Alarab, I., Cole, J., Ferreira, D., Johnson, B., Safdari, M., Dai, L., Arthornthurasuk, S., McAlister, I. C., Moyano, A. J., Pronin, A., Fan, J., Ramirez-Trinidad, A., Malysheva, Y., Pottmaier, D., Taheri, O., Stepanic, S., Perry, S., Askew, L., Rodríguez, R. A. H., Minissi, A. M. R., Lorena, R., Iyer, K., Fasiludeen, A. A., Clark, R., Ducey, J., Piza, M., Somrak, M., Vergo, E., Qin, J., Borbás, B., Chu, E., Lindsey, J., Jallon, A., McInnis, I. M. J., Chen, E., Semler, A., Gloor, L., Shah, T., Carauleanu, M., Lauer, P., Huy, T. D., Shahrtash, H., Duc, E., Lewark, L., Brown, A., Albanie, S., Weber, B., Vaz, W. S., Clavier, P., Fan, Y., Silva, G. P. R. e., Long, Lian, Abramovitch, M., Jiang, X., Mendoza, S., Islam, M., Gonzalez, J., Mavroudis, V., Xu, J., Kumar, P., Goswami, L. P., Bugas, D., Heydari, N., Jeanplong, F., Jansen, T., Pinto, A., Apronti, A., Galal, A., Ze-An, N., Singh, A., Jiang, T., Xavier, J. o. A., Agarwal, K. P., Berkani, M., Zhang, G., Du, Z., Junior, B. A. d. O., Malishev, D., Remy, N., Hartman, T. D., Tarver, T., Mensah, S., Loume, G. A., Morak, W., Habibi, F., Hoback, S., Cai, W., Gimenez, J., Montecillo, R. G., Łucki, J., Campbell, R., Sharma, A., Meer, K., Gul, S., Gonzalez, D. E., Alapont, X., Hoover, A., Chhablani, G., Vargus, F., Agarwal, A., Jiang, Y., Patil, D., Outevsky, D., Scaria, K. J., Maheshwari, R., Dendane, A., Shukla, P., Cartwright, A., Bogdanov, S., Mündler, N., Möller, S., Arnaboldi, L., Thaman, K., Siddiqi, M. R., Saxena, P., Gupta, H., Fruhauff, T., Sherman, G., Vincze, M., Usawasutsakorn, S., Ler, D., Radhakrishnan, A., Enyekwe,

I., Salauddin, S. M., Muzhen, J., Maksapetyan, A., Rossbach, V., Harjadi, C., Bahaloohoreh, M., Sparrow, C., Sidhu, J., Ali, S., Bian, S., Lai, J., Singer, E., Uro, J. L., Bateman, G., Sayed, M., Menshawy, A., Duclosel, D., Bezzi, D., Jain, Y., Aaron, A., Tiryakioglu, M., Siddh, S., Krenek, K., Shah, I. A., Jin, J., Creighton, S., Peskoff, D., EL-Wasif, Z., P, R., Richmond, M., McGowan, J., Patwardhan, T., Sun, H.-Y., Sun, T., Zubić, N., Sala, S., Ebert, S., Kaddour, J., Schottdorf, M., Wang, D., Petruzella, G., Meiburg, A., Medved, T., ElSheikh, A., Hebbar, S. A., Vaquero, L., Yang, X., Poulos, J., Zouhar, V., Bogdanik, S., Zhang, M., Sanz-Ros, J., Anugraha, D., Dai, Y., Nhu, A. N., Wang, X., Demircali, A. A., Jia, Z., Zhou, Y., Wu, J., He, M., Chandok, N., Sinha, A., Luo, G., Le, L., Noyé, M., Perełkiewicz, M., Pantidis, I., Qi, T., Purohit, S. S., Parcalabescu, L., Nguyen, T.-H., Winata, G. I., Ponti, E. M., Li, H., Dhole, K., Park, J., Abbondanza, D., Wang, Y., Nayak, A., Caetano, D. M., Wong, A. A. W. L., Rio-Chanona, M. d., Kondor, D., Francois, P., Chalstrey, E., Zsambok, J., Hoyer, D., Reddish, J., Hauser, J., Rodrigo-Ginés, F.-J., Datta, S., Shepherd, M., Kamphuis, T., Zhang, Q., Kim, H., Sun, R., Yao, J., Dernoncourt, F., Krishna, S., Rismanchian, S., Pu, B., Pinto, F., Wang, Y., Shridhar, K., Overholt, K. J., Briia, G., Nguyen, H., David, Bartomeu, S., Pang, T. C., Wecker, A., Xiong, Y., Li, F., Huber, L. S., Jaeger, J., Maddalena, R. D., Lù, X. H., Zhang, Y., Beger, C., Kon, P. T. J., Li, S., Sanker, V., Yin, M., Liang, Y., Zhang, X., Agrawal, A., Yifei, L. S., Zhang, Z., Cai, M., Sonmez, Y., Cozianu, C., Li, C., Slen, A., Yu, S., Park, H. K., Sarti, G., Briański, M., Stolfo, A., Nguyen, T. A., Zhang, M., Perlitz, Y., Hernandez-Orallo, J., Li, R., Shabani, A., Juefei-Xu, F., Dhingra, S., Zohar, O., Nguyen, M. C., Pondaven, A., Yilmaz, A., Zhao, X., Jin, C., Jiang, M., Todoran, S., Han, X., Kreuer, J., Rabern, B., Plassart, A., Maggetti, M., Yap, L., Geirhos, R., Kean, J., Wang, D., Mollaei, S., Sun, C., Yin, Y., Wang, S., Li, R., Chang, Y., Wei, A., Bizeul, A., Wang, X., Arrais, A. O., Mukherjee, K., Chamorro-Padial, J., Liu, J., Qu, X., Guan, J., Bouyamourn, A., Wu, S., Plomecka, M., Chen, J., Tang, M., Deng, J., Subramanian, S., Xi, H., Chen, H., Zhang, W., Ren, Y., Tu, H., Kim, S., Chen, Y., Marjanović, S. V., Ha, J., Luczyna, G., Ma, J. J., Shen, Z., Song, D., Zhang, C. E., Wang, Z., Gendron, G., Xiao, Y., Smucker, L., Weng, E., Lee, K. H., Ye, Z., Ermon, S., Lopez-Miguel, I. D., Knights, T., Gitter, A., Park, N., Wei, B., Chen, H., Pai, K., Elkhanany, A., Lin, H., Siedler, P. D., Fang, J., Mishra, R., Zsolnai-Fehér, K., Jiang, X., Khan, S., Yuan, J., Jain, R. K., Lin, X., Peterson, M., Wang, Z., Malusare, A., Tang, M., Gupta, I., Fosin, I., Kang, T., Dworakowska, B., Matsumoto, K., Zheng, G., Sewuster, G., Villanueva, J. P., Rannev, I., Chernyavsky, I., Chen, J., Banik, D., Racz, B., Dong, W., Wang, J., Bashmal, L., Gonçalves, D. V., Hu, W., Bar, K., Bohdal, O., Patlan, A. S., Dhuliawala, S., Geirhos, C., Wist, J., Kansal, Y., Chen, B., Tire, K., Yücel, A. T., Christof, B., Singla, V., Song, Z., Chen, S., Ge, J., Ponkshe, K., Park, I., Shi, T., Ma, M. Q., Mak, J., Lai, S., Moulin, A., Cheng, Z., Zhu, Z., Zhang, Z., Patil, V., Jha, K., Men, Q., Wu, J., Zhang, T., Vieira, B. H., Aji, A. F., Chung, J.-W., Mahfoud, M., Hoang, H. T., Sperzel, M., Hao, W., Meding, K., Xu, S., Kostakos, V., Manini, D., Liu, Y., Toukmaji, C., Paek, J., Yu, E., Demircali, A. E., Sun, Z., Dewerpe, I., Qin, H., Pflugfelder, R., Bailey, J., Morris, J., Heilala, V., Rosset, S., Yu, Z., Chen, P. E., Yeo, W., Jain, E., Yang, R., Chigurupati, S., Chernyavsky, J., Reddy, S. P., Venugopalan, S., Batra, H., Park, C. F., Tran, H., Maximiano, G., Zhang, G., Liang, Y., Shiyu, H., Xu, R., Pan, R., Suresh, S., Liu, Z., Gulati, S., Zhang, S., Turchin, P., Bartlett, C. W., Scotese, C. R., Cao, P. M., Wu, B., Karwowski, J., Scaramuzza, D., Nattanmai, A., McKellips, G., Cheraku, A., Suhail, A., Luo, E., Deng, M., Luo, J., Zhang, A., Jindel, K., Paek, J., Halevy, K., Baranov, A., Liu, M., Avadhanam, A., Zhang, D., Cheng, V., Ma, B., Fu, E., Do, L., Lass, J., Yang, H., Sunkari, S., Bharath, V., Ai, V., Leung, J., Agrawal, R., Zhou, A., Chen, K., Kalpathi, T., Xu, Z., Wang, G., Xiao, T., Maung, E., Lee, S., Yang, R., Yue, R., Zhao, B., Yoon, J., Sun, S., Singh, A., Luo, E., Peng, C., Osbey, T., Wang, T., Echeazu, D., Yang, H., Wu, T., Patel, S., Kulkarni, V., Sundarapandiyan, V., Zhang, A., Le, A., Nasim, Z., Yalam, S., Kasamsetty, R., Samal, S., Yang, H., Sun, D., Shah, N., Saha, A., Zhang, A., Nguyen, L., Nagumalli, L., Wang, K., Zhou, A., Wu, A., Luo, J., Telluri, A., Yue, S., Wang, A., and Hendrycks, D. Humanity's Last Exam, September 2025. URL http://arxiv.org/abs/2501.14249. arXiv:2501.14249 [cs].

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

Pimpale, G., Højmark, A., Scheurer, J., and Hobbhahn, M. Forecasting Frontier Language Model Agent Capabilities, March 2025. URL http://arxiv.org/abs/2502.15850. arXiv:2502.15850 [cs].

Polo, F. M., Somerstep, S., Choshen, L., Sun, Y., and Yurochkin, M. 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].

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A Graduate-Level Google-Proof Q&A Benchmark, November 2023. URL http://arxiv.org/abs/2311.12022. arXiv:2311.12022 [cs].

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

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

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. 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*, pp. 8732–8740. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6399.

Sandbrink, J., Hobbs, H., Swett, J., Dafoe, A., and Sandberg, A. Differential technology development: An innovation governance consideration for navigating technology risks, September 2022. URL https://papers.ssrn.com/abstract=4213670.

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

Scher, A. and Thiergart, L. 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.

Sevilla, J., Besiroglu, T., Cottier, B., You, J., Roldán, E., Villalobos, P., and Erdil, E. Can AI scaling continue through 2030? Technical report, Epoch AI, 2024. URL https://epoch.ai/blog/can-ai-scaling-continue-through-2030.

Shetty, M., Jain, N., Liu, J., Kethanaboyina, V., Sen, K., and Stoica, I. GSO: Challenging Software Optimization Tasks for Evaluating SWE-Agents, October 2025. URL http://arxiv.org/abs/2505.23671. arXiv:2505.23671 [cs].

Sirdeshmukh, V., Deshpande, K., Mols, J., Jin, L., Cardona, E.-Y., Lee, D., Kritz, J., Primack, W., Yue, S., and Xing, C. 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].

Srinivasa, R. S., Che, Z., Zhang, C. B. C., Mares, D., Hernandez, E., Park, J., Lee, D., Mangialardi, G., Ng, C., Cardona, E.-Y. H., Gunjal, A., He, Y., Liu, B., and Xing, C. TutorBench: A Benchmark To Assess Tutoring Capabilities Of Large Language Models, October 2025. URL http://arxiv.org/abs/2510.02663. arXiv:2510.02663 [cs].

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

Touzet, C., Papadatos, H., Murray, M., Quarks, O., Barrett, S., Boria, A. T., Perrier, E., Smith, M., and Campos, S. The Role of Risk Modeling in Advanced AI Risk Management, December 2025. URL http://arxiv.org/abs/2512.08723. arXiv:2512.08723 [cs].

Wang, C. J., Lee, D., Menghini, C., Mols, J., Doughty, J., Khoja, A., Lynch, J., Hendryx, S., Yue, S., and Hendrycks, D. EnigmaEval: A Benchmark of Long Multimodal Reasoning Challenges, February 2025. URL http://arxiv.org/abs/2502.08859. arXiv:2502.08859 [cs].

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *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.

Wasil, A. R., Barnett, P., Gerovitch, M., Hauksson, R., Reed, T., and Miller, J. W. 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.

Weij, T. v. d., Hofstätter, F., Jaffe, O., Brown, S. F., and Ward, F. R. AI Sandbagging: Language Models can Strategically Underperform on Evaluations, February 2025. URL http://arxiv.org/abs/2406.07358. arXiv:2406.07358 [cs].

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Dey, S., Shubh-Agrawal, Sandha, S. S., Naidu, S., Hegde, C., LeCun, Y., Goldstein, T., Neiswanger, W., and Goldblum, M. LiveBench: A Challenging, Contamination-Limited LLM Benchmark, April 2025. URL http://arxiv.org/abs/2406.19314. arXiv:2406.19314 [cs].

Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T. J., Cheng, Z., Shin, D., Lei, F., Liu, Y., Xu, Y., Zhou, S., Savarese, S., Xiong, C., Zhong, V., and Yu, T. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *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.

Xu, F. F., Song, Y., Li, B., Tang, Y., Jain, K., Bao, M., Wang, Z. Z., Zhou, X., Guo, Z., Cao, M., Yang, M., Lu, H. Y., Martin, A., Su, Z., Maben, L., Mehta, R., Chi, W., Jang, L., Xie, Y., Zhou, S., and Neubig, G. TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks, September 2025. URL http://arxiv.org/abs/2412.14161. arXiv:2412.14161 [cs].

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a Machine Really Finish Your Sentence? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proc. of ACL*, pp. 4791–4800, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Zhang, A. K., Perry, N., Dulepet, R., Ji, J., Menders, C., Lin, J. W., Jones, E., Hussein, G., Liu, S., Jasper, D., Peetathawatchai, P., Glenn, A., Sivashankar, V., Zamoshchin, D., Glikbarg, L., Askaryar, D., Yang, M., Zhang, T., Alluri, R., Tran, N., Sangpisit, R., Yiorkadjis, P., Osele, K., Raghupathi, G., Boneh, D., Ho, D. E., and Liang, P. 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].

Zhou, L., Pacchiardi, L., Martínez-Plumed, F., Collins, K. M., Moros-Daval, Y., Zhang, S., Zhao, Q., Huang, Y., Sun, L., Prunty, J. E., Li, Z., Sánchez-García, P., Chen, K. J., Casares, P. A. M., Zu, J., Burden, J., Mehrbakhsh, B., Stillwell, D., Cebrian, M., Wang, J., Henderson, P., Wu, S. T., Kyllonen, P. C., Cheke, L., Xie, X., and Hernández-Orallo, J. General Scales Unlock AI Evaluation with Explanatory and Predictive Power, March 2025. URL http://arxiv.org/abs/2503.06378. arXiv:2503.06378 [cs].

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. Retrodiction Analysis

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

B.1. Validation 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 for observations after the cutoff. We compare four model variants:

- Harvey hierarchical (joint hyperpriors across benchmarks)
- Harvey independent (separate fits per benchmark)
- Logistic hierarchical (joint hyperpriors)
- Logistic independent (separate fits)

Independent fits use identical model structure but estimate all parameters separately for each benchmark without shared hyperpriors.

B.2. Evaluation Metrics

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.

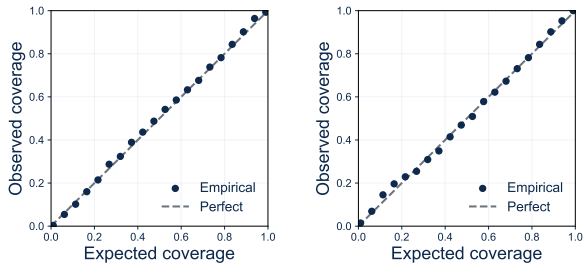
B.3. Results

Supp. Figure 5 and Supp. Table 1 summarize predictive performance.

Model	CRPS	RMSE
Harvey hierarchical	0.056	0.104
Harvey independent	0.054	0.102
Logistic hierarchical	0.053	0.100
Logistic independent	0.054	0.101

Table 1. Retrodiction performance (lower is better).

All four model variants achieve strong predictive performance, with CRPS values below 0.06 and good calibration, as shown in the calibration curve (Supp. Figure 5). The differences between models are small: variations in CRPS and



(a) Harvey Hierarchical *(b)* Harvey Independent

(c) Logistic Hierarchical *(d)* Logistic Independent

Figure 5. Calibration curves for the four models.

RMSE are comparable to the changes observed when adjusting the holdout date by a few data points, indicating that the choice between hierarchical versus independent structure, and between Harvey versus logistic functional forms, has a limited impact on prediction accuracy.

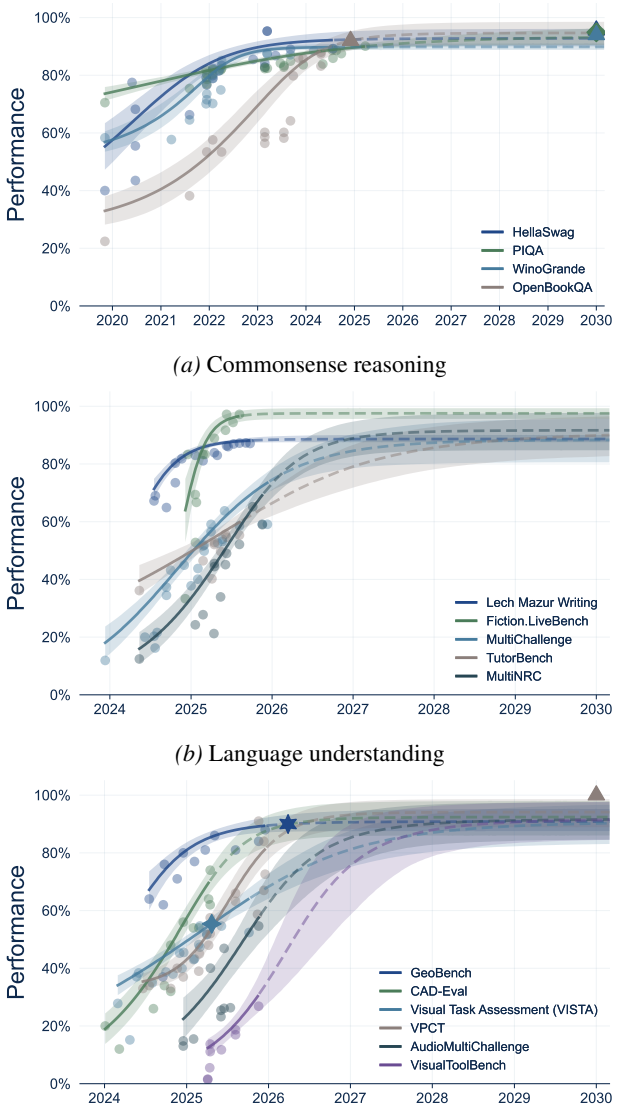
We select the Harvey hierarchical model for our main analysis. The hierarchical structure enables information sharing across benchmarks without degrading performance, which may particularly benefit predictions for data-sparse benchmarks. Although the Harvey model is more complex than the logistic, Bayesian 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.

C. 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.

D. Human Performance Baselines

Table 2 summarizes human performance baselines used in our analysis. We categorize human performance into four



(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.

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., Fields Medal mathematicians, 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).

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 & Hemang (2024)
Mathematical Reasoning			
MATH Level 5	Top Performer	90%	Hendrycks et al. (2021)
MATH Level 5	Skilled Generalist	40%	Hendrycks et al. (2021)
FrontierMath	Committee of Domain Experts	19%	Epoch AI (2025e)
FrontierMath	Committee of Domain Experts	35%	Epoch AI (2025e)
OTIS Mock AIME	High School Qualifier	53%	Chen (2025)
OTIS Mock AIME	High School Top Performer	93%	Chen (2025)
Core 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	Average Human	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 2. **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.