

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053
054

Position: The Window for Preventive Action Before Superhuman AI on Most Cognitive Tasks is Closing Rapidly

Anonymous Authors¹

Abstract

In this position paper, we argue that without coordinated intervention, AI systems will by default exceed human expert performance on most measurable 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 rou-

tinely 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,

¹Anonymous 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.

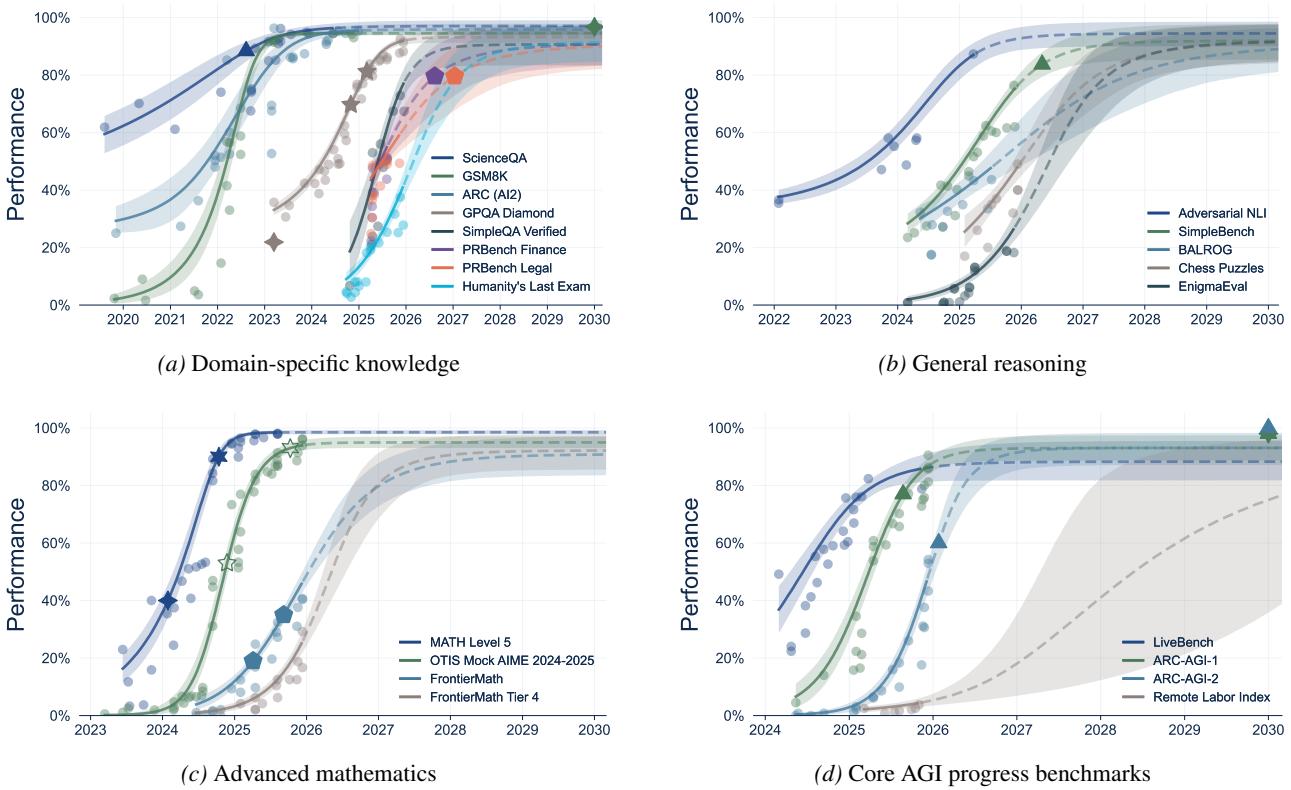


Figure 1. Benchmark trajectories across capability categories. Points show frontier model scores at release; solid lines show posterior median trajectories; shaded regions indicate 80% credible intervals. Dashed lines extrapolate to 2030. Human reference points are depicted by star-like symbols (representing 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 as-

sumption 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 sub-optimal 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 skewed-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-

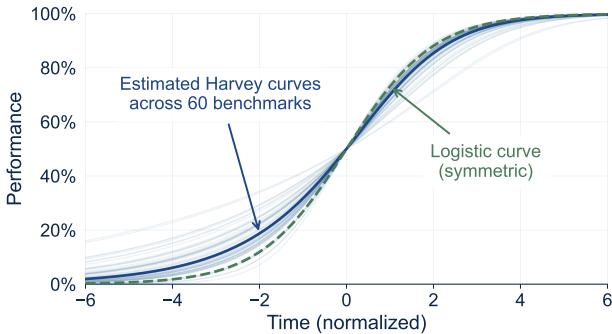


Figure 2. **Harvey curves capture asymmetric progress trajectories.** Fitted Harvey curves (blue) for 60 benchmarks show gradual acceleration followed by rapid saturation, compared to the symmetric logistic curve (green dashed). All curves are standardized to reach 50% performance at time 0 and have 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).

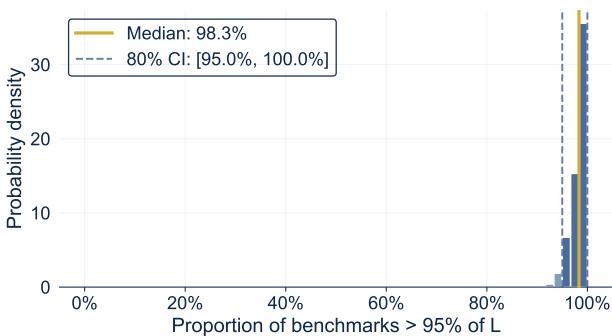


Figure 3. **Nearly all benchmarks projected to saturate by 2030.** Posterior distribution of the proportion of benchmarks reaching 95% of their maximum score range by 2030. The median is 98.3%, with 80% credible interval [95.0%, 100%], indicating that saturation of most benchmarks by 2030 is the default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

165
166
167
168
169
170 Remote Labor Index (RLI) measures completion rates
171 on real-world tasks from online freelance platforms.
172 LiveBench and ARC-AGI-2 should saturate soon; pre-
173 dictions for RLI are much more uncertain.

174
175 Benchmarks measuring capabilities with direct security im-
176 plications show particularly rapid trajectories (Figure 4).

- 177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
• **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, FictionLiveBench, 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 Assessemnt VISTA, VPCT,

44
45 Audio MultiChallenge, VisualToolBench) (ccmdi, 2026;
46 Epoch AI, 2025a; Scale AI, 2025; Brower, 2025; Gosai
47 et al., 2025; Guo et al., 2025).

48 Overall, all the projections we derived from the joint hier-
49 archical Harvey model estimate benchmark saturation by
50 2028 to 2030. Uncertainty intervals widen for longer fore-
51 cast horizons, but even the conservative (10th percentile)
52 projections place saturation for most benchmarks before
53 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.

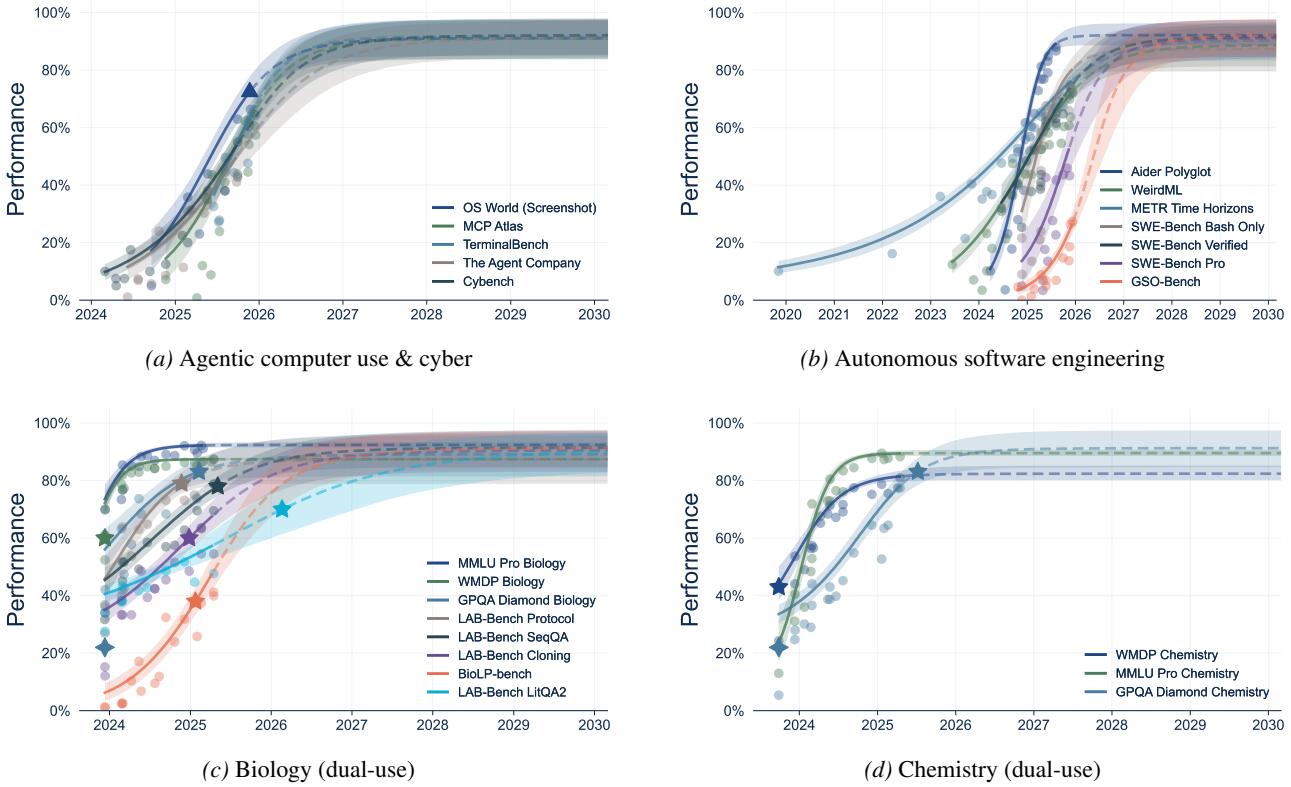


Figure 4. **Security-critical capability trajectories.** Benchmarks with direct safety implications show rapid progress, with most projected to saturate by 2027-2028. Starred markers indicate human expert baselines where available. On OS World, unfamiliar users achieve 72% (Xie et al., 2024); on biology 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-

330
 331
 332
 333
 334
 335
 336
 337
 338
 339
 340
 341
 342
 343
 344
 345
 346
 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.

- 385 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> [cs].
- 386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
15

- 440 URL <http://arxiv.org/abs/2110.14168>.
441 arXiv:2110.14168 [cs].
- 442
- 443 Dalrymple, D. d., Skalse, J., Bengio, Y., Russell, S.,
444 Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C.,
445 Goldhaber, B., Ammann, N., Abate, A., Halpern, J., Bar-
446 rett, C., Zhao, D., Zhi-Xuan, T., Wing, J., and Tenen-
447 baum, J. Towards Guaranteed Safe AI: A Framework for
448 Ensuring Robust and Reliable AI Systems, July 2024.
449 URL <http://arxiv.org/abs/2405.06624>.
450 arXiv:2405.06624 [cs].
- 451
- 452 Deng, X., Da, J., Pan, E., He, Y. Y., Ide, C., Garg, K.,
453 Lauffer, N., Park, A., Pasari, N., Rane, C., Sampath,
454 K., Krishnan, M., Kundurthy, S., Hendryx, S., Wang,
455 Z., Bharadwaj, V., Holm, J., Aluri, R., Zhang, C. B. C.,
456 Jacobson, N., Liu, B., and Kenstler, B. SWE-Bench Pro:
457 Can AI Agents Solve Long-Horizon Software Engineer-
458 ing Tasks?, November 2025. URL <http://arxiv.org/abs/2509.16941>. arXiv:2509.16941 [cs].
- 459
- 460 Dev, S., Teague, C., Ellison, G., Brady, K., Lee, Y.-C. J.,
461 Gebauer, S. L., Bradley, H. A., Maciorowski, D., Per-
462 saud, B., Despanie, J., Del Castello, B., Worland, A.,
463 Miller, M., Salas, A., Nguyen, D., Liu, J., Johnson, J.,
464 Sloan, A., Stonehouse, W., Merrill, T., Goode, T., McK-
465 elvey, G., and Guest, E. Toward Comprehensive Bench-
466 marking of the Biological Knowledge of Frontier Large
467 Language Models. Technical report, RAND Corpora-
468 tion, 2025. URL https://www.rand.org/pubs/research_reports/RRA3797-1.html.
- 469
- 470
- 471 Epoch AI. AI Benchmarking Hub, November 2024a. URL
472 <https://epoch.ai/benchmarks/use-thi>s-data.
- 473
- 474
- 475 Epoch AI. SWE-bench Verified, 2024b. URL <https://epoch.ai/benchmarks/swe-bench-verif>ied.
- 476
- 477
- 478
- 479 Epoch AI. CadEval, 2025a. URL <https://epoch.ai/benchmarks/cad-eval>.
- 480
- 481
- 482 Epoch AI. Chess Puzzles, December 2025b. URL <https://epoch.ai/benchmarks/chess-puzzles>.
- 483
- 484
- 485 Epoch AI. Data on Frontier AI Data Centers, September
486 2025c. URL <https://epoch.ai/data/data-centers>.
- 487
- 488
- 489 Epoch AI. Fiction.liveBench, February 2025d. URL <https://epoch.ai/benchmarks/fictionlivebench>.
- 490
- 491
- 492 Epoch AI. FrontierMath, 2025e. URL <https://epoch.ai/benchmarks/frontiermath>.
- 493
- 494
- 495
- 496
- 497
- 498
- 499
- 500 Epoch AI. OTIS Mock AIME 2024-2025, 2025f. URL
501 <https://epoch.ai/benchmarks/otis-mock-aime-2024-2025>.
- 502
- 503 Erdil, E. and Besiroglu, T. Algorithmic progress in com-
504 puter vision, August 2023. URL <http://arxiv.org/abs/2212.05153>. arXiv:2212.05153 [cs].
- 505
- 506 Fabbri, A. R., Mares, D., Flores, J., Mankikar, M., Hernan-
507 dez, E., Lee, D., Liu, B., and Xing, C. MultiNRC: A
508 Challenging and Native Multilingual Reasoning Evalu-
509 ation Benchmark for LLMs, July 2025. URL <http://arxiv.org/abs/2507.17476>. arXiv:2507.17476 [cs].
- 510
- 511 Gosai, A., Vuong, T., Tyagi, U., Li, S., You, W., Bavare,
512 M., Uçar, A., Fang, Z., Jang, B., Liu, B., and He, Y. Au-
513 dio MultiChallenge: A Multi-Turn Evaluation of Spoken
514 Dialogue Systems on Natural Human Interaction, De-
515 cember 2025. URL <http://arxiv.org/abs/2512.14865>. arXiv:2512.14865 [cs].
- 516
- 517 Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O.
518 When Will AI Exceed Human Performance? Evidence
519 from AI Experts, May 2018. URL <http://arxiv.org/abs/1705.08807>. arXiv:1705.08807 [cs].
- 520
- 521 Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S.,
522 Weinstein-Raun, B., Brauner, J., and Korzekwa, R. C.
523 Thousands of AI Authors on the Future of AI. *jair*, 84,
524 October 2025. ISSN 1076-9757. doi: 10.1613/jair.1.1
525 9087. URL <http://arxiv.org/abs/2401.02843>. arXiv:2401.02843 [cs].
- 526
- 527 Guo, X., Tyagi, U., Gosai, A., Vergara, P., Park, J.,
528 Montoya, E. G. H., Zhang, C. B. C., Hu, B., He, Y.,
529 Liu, B., and Srinivasa, R. S. Beyond Seeing: Evalu-
530 ating Multimodal LLMs on Tool-Enabled Image Per-
531 ception, Transformation, and Reasoning, October 2025.
532 URL <http://arxiv.org/abs/2510.12712>. arXiv:2510.12712 [cs].
- 533
- 534 Haas, L., Yona, G., D'Antonio, G., Goldshtain, S., and
535 Das, D. SimpleQA Verified: A Reliable Factuality
536 Benchmark to Measure Parametric Knowledge, Septem-
537 ber 2025. URL <http://arxiv.org/abs/2509.07968>. arXiv:2509.07968 [cs].
- 538
- 539 Harvey, A. C. Time Series Forecasting Based on the
540 Logistic Curve. *Journal of the Operational Research
541 Society*, 35(7):641–646, July 1984. ISSN 0160-5682.
542 doi: 10.1057/jors.1984.128. URL <https://doi.org/10.1057/jors.1984.128>. eprint:
543 <https://doi.org/10.1057/jors.1984.128>.
- 544
- 545 Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart,
546 S., Tang, E., Song, D., and Steinhardt, J. Measuring
- 547

- 495 Mathematical Problem Solving With the MATH Dataset,
496 November 2021. URL <http://arxiv.org/abs/2103.03874>. arXiv:2103.03874 [cs].
497
- 498 Ho, A., Denain, J.-S., Atanasov, D., Albanie, S., and Shah,
499 R. A Rosetta Stone for AI Benchmarks, November 2025.
500 URL <http://arxiv.org/abs/2512.00193>. arXiv:2512.00193 [cs].
501
- 502 Ihle, H. T. WeirdML, 2025. URL <https://htihle.gitub.io/weirdml.html>.
- 503 Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press,
504 O., and Narasimhan, K. R. SWE-bench: Can Language
505 Models Resolve Real-world Github Issues? In *Proc. of
506 ICLR*. OpenReview.net, 2024. URL [https://open
507 review.net/forum?id=VTF8yNQM66](https://openreview.net/forum?id=VTF8yNQM66).
- 508 Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B.,
509 Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and
510 Amodei, D. Scaling Laws for Neural Language Models,
511 January 2020. URL <http://arxiv.org/abs/2001.08361>. arXiv:2001.08361 [cs].
512
- 513 Kipnis, A., Voudouris, K., Buschoff, L. M. S., and Schulz,
514 E. metabench – A Sparse Benchmark of Reasoning and
515 Knowledge in Large Language Models, February 2025.
516 URL <http://arxiv.org/abs/2407.12844>. arXiv:2407.12844 [cs].
517
- 518 Kučinskas, S., Rosenberg, J., de Castro, R. C., Jacobs, Z.,
519 Canedy, J., Tetlock, P. E., and Karger, E. Assessing
520 Near-Term Accuracy in the Existential Risk Persuasion
521 Tournament. Technical report, Forecasting Research
522 Institute, September 2025. URL <https://forecastingresearch.org/near-term-xpt-accuracy>.
- 523 Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin,
524 M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V.,
525 Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic,
526 N., Miles, L. H., Nix, S., Lin, T., Parikh, N., Rein, D.,
527 Sato, L. J. K., Wijk, H., Ziegler, D. M., Barnes, E., and
528 Chan, L. Measuring AI Ability to Complete Long Tasks,
529 March 2025. URL <http://arxiv.org/abs/2503.14499>. arXiv:2503.14499 [cs].
530
- 531 Laurent, J. M., Janizek, J. D., Ruzo, M., Hinks, M. M.,
532 Hammerling, M. J., Narayanan, S., Ponnappati, M.,
533 White, A. D., and Rodrigues, S. G. LAB-Bench: Mea-
534 suring Capabilities of Language Models for Biology Re-
535 search, July 2024. URL <http://arxiv.org/abs/2407.10362>. arXiv:2407.10362 [cs].
536
- 537 Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti,
538 A., Li, J. D., Dombrowski, A.-K., Goel, S., Phan, L.,
539 Mukobi, G., Helm-Burger, N., Lababidi, R., Justen, L.,
540
- 541 Liu, A. B., Chen, M., Barrass, I., Zhang, O., Zhu,
542 X., Tamirisa, R., Bharathi, B., Khoja, A., Zhao, Z.,
543 Herbert-Voss, A., Breuer, C. B., Marks, S., Patel, O.,
544 Zou, A., Mazeika, M., Wang, Z., Oswal, P., Lin, W.,
545 Hunt, A. A., Tienken-Harder, J., Shih, K. Y., Talley,
546 K., Guan, J., Kaplan, R., Steneker, I., Campbell, D.,
547 Jokubaitis, B., Levinson, A., Wang, J., Qian, W., Kar-
548 makar, K. K., Basart, S., Fitz, S., Levine, M., Ku-
549 maraguru, P., Tupakula, U., Varadharajan, V., Wang, R.,
550 Shoshtaishvili, Y., Ba, J., Esveld, K. M., Wang, A., and
551 Hendrycks, D. The WMDP Benchmark: Measuring and
552 Reducing Malicious Use With Unlearning, May 2024a.
553 URL <http://arxiv.org/abs/2403.03218>. arXiv:2403.03218 [cs].
554
- 555 Li, Q., Cui, L., Zhao, X., Kong, L., and Bi, W. GSM-
556 Plus: A Comprehensive Benchmark for Evaluating the
557 Robustness of LLMs as Mathematical Problem Solvers,
558 July 2024b. URL <http://arxiv.org/abs/2402.19255>. arXiv:2402.19255 [cs].
559
- 560 Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu,
561 S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn
562 to Explain: Multimodal Reasoning via Thought Chains
563 for Science Question Answering, October 2022. URL
564 <http://arxiv.org/abs/2209.09513>. arXiv:2209.09513 [cs].
565
- 566 Maier, A., Maier, A., and David, T. Take Goodhart Se-
567 riously: Principled Limit on General-Purpose AI Opti-
568 mization, October 2025. URL <http://arxiv.org/abs/2510.02840>. arXiv:2510.02840 [cs].
569
- 570 Mazeika, M., Gatti, A., Menghini, C., Sehwag, U. M.,
571 Singhal, S., Orlovskiy, Y., Basart, S., Sharma, M.,
572 Peskoff, D., Lau, E., Lim, J., Carroll, L., Blair, A.,
573 Sivakumar, V., Basu, S., Kenstler, B., Ma, Y., Michael,
574 J., Li, X., Ingebretsen, O., Mehta, A., Mottola, J., Teich-
575 mann, J., Yu, K., Shaik, Z., Khoja, A., Ren, R., Hausen-
576 loy, J., Phan, L., Htet, Y., Aich, A., Rabbani, T., Shah,
577 V., Novykov, A., Binder, F., Chugunov, K., Ramirez, L.,
578 Geralnik, M., Mesura, H., Lee, D., Cardona, E.-Y. H.,
579 Diamond, A., Yue, S., Wang, A., Liu, B., Hernandez,
580 E., and Hendrycks, D. Remote Labor Index: Mea-
581 suring AI Automation of Remote Work, October 2025.
582 URL <http://arxiv.org/abs/2510.26787>. arXiv:2510.26787 [cs].
583
- 584 Mazur, L. lechmazur/writing, January 2026. URL <https://github.com/lechmazur/writing>. original-
585 date: 2025-01-05T08:48:04Z.
- 586 Merrill, M. A., Shaw, A. G., Carlini, N., Li, B., Raj, H.,
587 Bercovich, I., Shi, L., Shin, J. Y., Walshe, T., Buchanan,
588 E. K., Shen, J., Ye, G., Lin, H., Poulos, J., Wang,
589 M., Nezhurina, M., Jitsev, J., Lu, D., Mastromichalakis,

- 550 O. M., Xu, Z., Chen, Z., Liu, Y., Zhang, R., Chen,
 551 L. L., Kashyap, A., Uslu, J.-L., Li, J., Wu, J., Yan, M.,
 552 Bian, S., Sharma, V., Sun, K., Dillmann, S., Anand,
 553 A., Lanpouthakoun, A., Koopah, B., Hu, C., Guha, E.,
 554 Dreiman, G. H. S., Zhu, J., Krauth, K., Zhong, L.,
 555 Muennighoff, N., Amanfu, R., Tan, S., Pimpalgaonkar,
 556 S., Aggarwal, T., Lin, X., Lan, X., Zhao, X., Liang,
 557 Y., Wang, Y., Wang, Z., Zhou, C., Heineman, D., Liu,
 558 H., Trivedi, H., Yang, J., Lin, J., Shetty, M., Yang,
 559 M., Omi, N., Raoof, N., Li, S., Zhuo, T. Y., Lin, W.,
 560 Dai, Y., Wang, Y., Chai, W., Zhou, S., Wahdany, D.,
 561 She, Z., Hu, J., Dong, Z., Zhu, Y., Cui, S., Saiyed,
 562 A., Kolbeinsson, A., Hu, J., Rytting, C. M., Marten, R.,
 563 Wang, Y., Dimakis, A., Konwinski, A., and Schmidt, L.
 564 Terminal-Bench: Benchmarking Agents on Hard, Realistic
 565 Tasks in Command Line Interfaces, January 2026.
 566 URL <http://arxiv.org/abs/2601.11868>.
 567 arXiv:2601.11868 [cs].
- 568 Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can
 569 a Suit of Armor Conduct Electricity? A New Dataset
 570 for Open Book Question Answering, September 2018.
 571 URL <http://arxiv.org/abs/1809.02789>.
 572 arXiv:1809.02789 [cs].
- 573 Murray, M., Barrett, S., Papadatos, H., Quarks, O., Smith,
 574 M., Boria, A. T., Touzet, C., and Campos, S. A Methodology
 575 for Quantitative AI Risk Modeling, December 2025.
 576 URL <http://arxiv.org/abs/2512.08844>.
 577 arXiv:2512.08844 [cs].
- 578 Ngo, R., Chan, L., and Mindermann, S. The Alignment
 579 Problem from a Deep Learning Perspective, May 2025.
 580 URL <http://arxiv.org/abs/2209.00626>.
 581 arXiv:2209.00626 [cs].
- 582 Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J.,
 583 and Kiela, D. Adversarial NLI: A New Benchmark for
 584 Natural Language Understanding. In Jurafsky, D., Chai,
 585 J., Schluter, N., and Tetreault, J. (eds.), *Proc. of ACL*,
 586 pp. 4885–4901, Online, 2020. Association for Computational
 587 Linguistics. doi: 10.18653/v1/2020.acl-main.441.
 588 URL <https://aclanthology.org/2020.acl-main.441>.
- 589 OpenAI. Learning to Reason with LLMs, September 2024.
 590 URL <https://openai.com/index/learning-to-reason-with-langs>.
- 591 Paglieri, D., Cupial, B., Coward, S., Piterbarg, U., Wol-
 592 czczyk, M., Khan, A., Pignatelli, E., Kuciński, L., Pinto,
 593 L., Fergus, R., Foerster, J. N., Parker-Holder, J., and
 594 Rocktäschel, T. BALROG: Benchmarking Agentic
 595 LLM and VLM Reasoning On Games, April 2025.
 596 URL <http://arxiv.org/abs/2411.13543>.
 597 arXiv:2411.13543 [cs].
- 598 Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H.,
 599 Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi,
 600 M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren,
 601 R., Hausenloy, J., Zhang, O., Mazeika, M., Dodonov,
 602 D., Nguyen, T., Lee, J., Anderson, D., Doroshenko, M.,
 603 Stokes, A. C., Mahmood, M., Pokutnyi, O., Iskra, O.,
 604 Wang, J. P., Levin, J.-C., Kazakov, M., Feng, F., Feng,
 605 S. Y., Zhao, H., Yu, M., Gangal, V., Zou, C., Wang, Z.,
 606 Popov, S., Gerbicz, R., Galgon, G., Schmitt, J., Yeadon,
 607 W., Lee, Y., Sauers, S., Sanchez, A., Giska, F., Roth,
 608 M., Riis, S., Utpala, S., Burns, N., Goshu, G. M., Naiya,
 609 M. M., Agu, C., Giboney, Z., Cheatom, A., Fournier-
 610 Facio, F., Crowson, S.-J., Finke, L., Cheng, Z., Zam-
 611 pese, J., Hoerr, R. G., Nandor, M., Park, H., Gehrunger,
 612 T., Cai, J., McCarty, B., Garretson, A. C., Taylor, E.,
 613 Sileo, D., Ren, Q., Qazi, U., Li, L., Nam, J., Wydallis,
 614 J. B., Arkhipov, P., Shi, J. W. L., Bacho, A., Willcocks,
 615 C. G., Cao, H., Motwani, S., Santos, E. d. O., Veith, J.,
 616 Vendrow, E., Cojoc, D., Zenitani, K., Robinson, J., Tang,
 617 L., Li, Y., Vendrow, J., Fraga, N. W., Kuchkin, V., Mak-
 618 simov, A. P., Marion, P., Efremov, D., Lynch, J., Liang,
 619 K., Mikov, A., Gritsevskiy, A., Guillod, J., Demir, G.,
 620 Martinez, D., Pageler, B., Zhou, K., Soori, S., Press, O.,
 621 Tang, H., Rissone, P., Green, S. R., Brüssel, L., Twanya,
 622 M., Dieuleveut, A., Imperial, J. M., Prabhu, A., Yang,
 623 J., Crispino, N., Rao, A., Zvonkine, D., Loiseau, G.,
 624 Kalinin, M., Lukas, M., Manolescu, C., Stambaugh,
 625 N., Mishra, S., Hogg, T., Bosio, C., Coppola, B. P.,
 626 Salazar, J., Jin, J., Sayous, R., Ivanov, S., Schwaller,
 627 P., Senthilkuma, S., Bran, A. M., Algaba, A., Houte, K.,
 628 V. d., Sypt, L. V. D., Verbeken, B., Noever, D., Kopy-
 629 lov, A., Myklebust, B., Li, B., Schut, L., Zhettonozhskii,
 630 E., Yuan, Q., Lim, D., Stanley, R., Yang, T., Maar, J.,
 631 Wykowski, J., Oller, M., Sahu, A., Ardito, C. G., Hu, Y.,
 632 Kamdoum, A. G. K., Jin, A., Vilchis, T. G., Zu, Y., Lack-
 633 ner, M., Koppel, J., Sun, G., Antonenko, D. S., Chern, S.,
 634 Zhao, B., Arsene, P., Cavanagh, J. M., Li, D., Shen, J.,
 635 Crisostomi, D., Zhang, W., Dehghan, A., Ivanov, S., Per-
 636 rella, D., Kaparov, N., Zang, A., Sucholutsky, I., Khar-
 637 lamova, A., Orel, D., Poritski, V., Ben-David, S., Berger,
 638 Z., Whitfill, P., Foster, M., Munro, D., Ho, L., Sivarajan,
 639 S., Hava, D. B., Kuchkin, A., Holmes, D., Rodriguez-
 640 Romero, A., Sommerhage, F., Zhang, A., Moat, R.,
 641 Schneider, K., Kazibwe, Z., Clarke, D., Kim, D. H.,
 642 Dias, F. M., Fish, S., Elser, V., Kreiman, T., Vilchis, V.,
 643 E. G., Klose, I., Anantheswaran, U., Zweiger, A., Rawal,
 644 K., Li, J., Nguyen, J., Daans, N., Heidinger, H., Rad-
 645 dionov, M., Rozhoň, V., Ginis, V., Stump, C., Cohen,
 646 N., Pościata, R., Tkadlec, J., Goldfarb, A., Wang, C.,
 647 Padlewski, P., Barzowski, S., Montgomery, K., Stendall,
 648 R., Tucker-Foltz, J., Stade, J., Rogers, T. R., Goertzen,
 649 T., Grabb, D., Shukla, A., Givré, A., Ambay, J. A., Sen,
 650 A., Aziz, M. F., Inlow, M. H., He, H., Zhang, L., Kad-
 651 dar, Y., Ängquist, I., Chen, Y., Wang, H. K., Ramakrish-

605 nan, K., Thornley, E., Terpin, A., Schoelkopf, H., Zheng,
 606 E., Carmi, A., Brown, E. D. L., Zhu, K., Bartolo, M.,
 607 Wheeler, R., Stehberger, M., Bradshaw, P., Heimonen,
 608 J. P., Sridhar, K., Akov, I., Sandlin, J., Makarychev, Y.,
 609 Tam, J., Hoang, H., Cunningham, D. M., Goryachev,
 610 V., Patramanis, D., Krause, M., Redenti, A., Aldous,
 611 D., Lai, J., Coleman, S., Xu, J., Lee, S., Magoulas, I.,
 612 Zhao, S., Tang, N., Cohen, M. K., Paradise, O., Kirchner,
 613 J. H., Ovchinnikov, M., Matos, J. O., Shenoy, A.,
 614 Wang, M., Nie, Y., Sztyber-Betley, A., Faraboschi, P.,
 615 Riblet, R., Crozier, J., Halasyamani, S., Verma, S., Joshi,
 616 P., Meril, E., Ma, Z., Andréoletti, J., Singh, R., Plat-
 617 nick, J., Nevirkovets, V., Basler, L., Ivanov, A., Khoury,
 618 S., Gustafsson, N., Piccardo, M., Mostaghimi, H., Chen,
 619 Q., Singh, V., Khanh, T. Q., Rosu, P., Szlyk, H., Brown,
 620 Z., Narayan, H., Menezes, A., Roberts, J., Alley, W.,
 621 Sun, K., Patel, A., Lamparth, M., Reuel, A., Xin, L., Xu,
 622 H., Loader, J., Martin, F., Wang, Z., Achilleos, A., Preu,
 623 T., Korbak, T., Bosio, I., Kazemi, F., Chen, Z., Bálint,
 624 B., Lo, E. J. Y., Wang, J., Nunes, M. I. S., Milbauer,
 625 J., Bari, M. S., Wang, Z., Ansarinejad, B., Sun, Y., Du-
 626 rand, S., Elgnainy, H., Douville, G., Tordera, D., Bal-
 627 abanian, G., Wolff, H., Kvistad, L., Milliron, H., Sakor,
 628 A., Eron, M., O. A. F. D., Shah, S., Zhou, X., Kamalov,
 629 F., Abdoli, S., Santens, T., Barkan, S., Tee, A., Zhang,
 630 R., Tomasiello, A., Luca, G. B. D., Looi, S.-Z., Le, V.-
 631 K., Kolt, N., Pan, J., Rodman, E., Drori, J., Fossum,
 632 C. J., Muennighoff, N., Jagota, M., Pradeep, R., Fan, H.,
 633 Eicher, J., Chen, M., Thaman, K., Merrill, W., Firsching,
 634 M., Harris, C., Ciobăcă, S., Gross, J., Pandey, R., Gusev,
 635 I., Jones, A., Agnihotri, S., Zhelnov, P., Mofayzei, M.,
 636 Piperski, A., Zhang, D. K., Dobarskyi, K., Leventov, R.,
 637 Soroko, I., Duersch, J., Taamazyan, V., Ho, A., Ma, W.,
 638 Held, W., Xian, R., Zebaze, A. R., Mohamed, M., Leser,
 639 J. N., Yuan, M. X., Yacar, L., Lengler, J., Olszewska, K.,
 640 Fratta, C. D., Oliveira, E., Jackson, J. W., Zou, A., Chi-
 641 dambaram, M., Manik, T., Haffenden, H., Stander, D.,
 642 Dasouqi, A., Shen, A., Golshani, B., Stap, D., Kretov,
 643 E., Uzhou, M., Zhidkovskaya, A. B., Winter, N., Ro-
 644 driguez, M. O., Lauff, R., Wehr, D., Tang, C., Hossain,
 645 Z., Phillips, S., Samuele, F., Ekström, F., Hammon, A.,
 646 Patel, O., Farhidi, F., Medley, G., Mohammadzadeh, F.,
 647 Peñaflor, M., Kassahun, H., Friedrich, A., Perez, R. H.,
 648 Pyda, D., Sakal, T., Dhamane, O., Mirabadi, A. K., Hall-
 649 man, E., Okutsu, K., Battaglia, M., Maghsoudimehrabani,
 650 M., Amit, A., Hulbert, D., Pereira, R., Weber, S.,
 651 Handoko, Peristy, A., Malina, S., Mehkary, M., Aly,
 652 R., Reidegeld, F., Dick, A.-K., Friday, C., Singh, M.,
 653 Shapourian, H., Kim, W., Costa, M., Gurdogan, H., Ku-
 654 mar, H., Ceconello, C., Zhuang, C., Park, H., Carroll,
 655 M., Tawfeek, A. R., Steinerberger, S., Aggarwal, D.,
 656 Kirchhof, M., Dai, L., Kim, E., Ferret, J., Shah, J., Wang,
 657 Y., Yan, M., Burdzy, K., Zhang, L., Franca, A., Pham,
 658 D. T., Loh, K. Y., Robinson, J., Jackson, A., Giordano,

659 P., Petersen, P., Cosma, A., Colino, J., White, C., Votava,
 660 J., Vinnikov, V., Delaney, E., Spelda, P., Stritecky, V.,
 661 Shahid, S. M., Mourrat, J.-C., Vetoshkin, L., Sponselee,
 662 K., Bacho, R., Yong, Z.-X., Rosa, F. d. I., Cho, N., Li,
 663 X., Malod, G., Weller, O., Albani, G., Lang, L., Lau-
 664 rendeau, J., Kazakov, D., Adesanya, F., Portier, J., Hol-
 665 lom, L., Souza, V., Zhou, Y. A., Degorre, J., Yalın, Y.,
 666 Obikoya, G. D., Rai, Bigi, F., Boscá, M. C., Shumar, O.,
 667 Bacho, K., Recchia, G., Popescu, M., Shulga, N., Tan-
 668 wie, N. M., Lux, T. C. H., Rank, B., Ni, C., Brooks, M.,
 669 Yakimchyk, A., Huanxu, Liu, Cavalleri, S., Häggström,
 670 O., Verkama, E., Newbould, J., Gundlach, H., Brito-
 671 Santana, L., Amaro, B., Vajipey, V., Grover, R., Wang,
 672 T., Kratish, Y., Li, W.-D., Gopi, S., Caciolai, A., Witt,
 673 C. S. d., Hernández-Cámara, P., Rodolà, E., Robins, J.,
 674 Williamson, D., Cheng, V., Raynor, B., Qi, H., Segev,
 675 B., Fan, J., Martinson, S., Wang, E. Y., Hausknecht,
 676 K., Brenner, M. P., Mao, M., Demian, C., Kassani, P.,
 677 Zhang, X., Avagian, D., Scipio, E. J., Ragoler, A., Tan,
 678 J., Sims, B., Plecnik, R., Kirtland, A., Bodur, O. F.,
 679 Shinde, D. P., Labrador, Y. C. L., Adoul, Z., Zekry, M.,
 680 Karakoc, A., Santos, T. C. B., Shamseldeen, S., Karim,
 681 L., Liakhovitskaia, A., Resman, N., Farina, N., Gonza-
 682 lez, J. C., Maayan, G., Anderson, E., Pena, R. D. O., Kel-
 683 ley, E., Mariji, H., Pouriamanesh, R., Wu, W., Finocchio,
 684 R., Alarab, I., Cole, J., Ferreira, D., Johnson, B., Saf-
 685 dari, M., Dai, L., Arthonthurasuk, S., McAlister, I. C.,
 686 Moyano, A. J., Pronin, A., Fan, J., Ramirez-Trinidad,
 687 A., Malysheva, Y., Pottmaier, D., Taheri, O., Stepanic,
 688 S., Perry, S., Askew, L., Rodríguez, R. A. H., Minissi,
 689 A. M. R., Lorena, R., Iyer, K., Fasiludeen, A. A., Clark,
 690 R., Ducey, J., Piza, M., Somrak, M., Vergo, E., Qin,
 691 J., Borbás, B., Chu, E., Lindsey, J., Jallon, A., McIn-
 692 nnis, I. M. J., Chen, E., Semler, A., Gloor, L., Shah, T.,
 693 Carauleanu, M., Lauer, P., Huy, T. D., Shahrtash, H.,
 694 Duc, E., Lewark, L., Brown, A., Albanie, S., Weber,
 695 B., Vaz, W. S., Clavier, P., Fan, Y., Silva, G. P. R. e.,
 696 Long, Lian, Abramovitch, M., Jiang, X., Mendoza, S.,
 697 Islam, M., Gonzalez, J., Mavroudis, V., Xu, J., Kumar,
 698 P., Goswami, L. P., Bugas, D., Heydari, N., Jeanplong,
 699 F., Jansen, T., Pinto, A., Apronti, A., Galal, A., Ze-An,
 700 N., Singh, A., Jiang, T., Xavier, J. o. A., Agarwal, K. P.,
 701 Berkani, M., Zhang, G., Du, Z., Junior, B. A. d. O., Mal-
 702 ishev, D., Remy, N., Hartman, T. D., Tarver, T., Men-
 703 sah, S., Loume, G. A., Morak, W., Habibi, F., Hoback,
 704 S., Cai, W., Gimenez, J., Montecillo, R. G., Łucki, J.,
 705 Campbell, R., Sharma, A., Meer, K., Gul, S., Gonza-
 706 lez, D. E., Alapont, X., Hoover, A., Chhablani, G., Var-
 707 gus, F., Agarwal, A., Jiang, Y., Patil, D., Outevsky, D.,
 708 Scaria, K. J., Maheshwari, R., Dendane, A., Shukla, P.,
 709 Cartwright, A., Bogdanov, S., Mündler, N., Möller, S.,
 710 Arnaboldi, L., Thaman, K., Siddiqi, M. R., Saxena, P.,
 711 Gupta, H., Fruhauff, T., Sherman, G., Vincze, M., Us-
 712 awasutsakorn, S., Ler, D., Radhakrishnan, A., Enyekwe,

- 660 I., Salauddin, S. M., Muzhen, J., Maksapetyan, A., Ross-
661 bach, V., Harjadi, C., Bahalooohoreh, M., Sparrow, C.,
662 Sidhu, J., Ali, S., Bian, S., Lai, J., Singer, E., Uro, J. L.,
663 Bateman, G., Sayed, M., Menshawy, A., Duclosel, D.,
664 Bezzi, D., Jain, Y., Aaron, A., Tiryakioglu, M., Siddh,
665 S., Krenek, K., Shah, I. A., Jin, J., Creighton, S., Peskoff,
666 D., EL-Wasif, Z., P. R., Richmond, M., McGowan, J.,
667 Patwardhan, T., Sun, H.-Y., Sun, T., Zubić, N., Sala,
668 S., Ebert, S., Kaddour, J., Schottdorf, M., Wang, D.,
669 Petruzella, G., Meiburg, A., Medved, T., ElSheikh, A.,
670 Hebbar, S. A., Vaquero, L., Yang, X., Poulos, J., Zouhar,
671 V., Bogdanik, S., Zhang, M., Sanz-Ros, J., Anugraha,
672 D., Dai, Y., Nhu, A. N., Wang, X., Demircali, A. A., Jia,
673 Z., Zhou, Y., Wu, J., He, M., Chandok, N., Sinha, A.,
674 Luo, G., Le, L., Noyé, M., Perelkiewicz, M., Pantidis, I.,
675 Qi, T., Purohit, S. S., Parcalabescu, L., Nguyen, T.-H.,
676 Winata, G. I., Ponti, E. M., Li, H., Dhole, K., Park, J.,
677 Abbondanza, D., Wang, Y., Nayak, A., Caetano, D. M.,
678 Wong, A. A. W. L., Rio-Chanona, M. d., Kondor, D.,
679 Francois, P., Chalstrey, E., Zsambok, J., Hoyer, D., Red-
680 dish, J., Hauser, J., Rodrigo-Ginés, F.-J., Datta, S., Shep-
681 herd, M., Kamphuis, T., Zhang, Q., Kim, H., Sun, R.,
682 Yao, J., Dernoncourt, F., Krishna, S., Rismanchian, S.,
683 Pu, B., Pinto, F., Wang, Y., Shridhar, K., Overholt, K. J.,
684 Briia, G., Nguyen, H., David, Bartomeu, S., Pang, T. C.,
685 Wecker, A., Xiong, Y., Li, F., Huber, L. S., Jaeger, J.,
686 Maddalena, R. D., Lù, X. H., Zhang, Y., Beger, C., Kon,
687 P. T. J., Li, S., Sanker, V., Yin, M., Liang, Y., Zhang,
688 X., Agrawal, A., Yifei, L. S., Zhang, Z., Cai, M., Son-
689 mez, Y., Cozianu, C., Li, C., Slen, A., Yu, S., Park, H. K.,
690 Sarti, G., Briański, M., Stolfo, A., Nguyen, T. A., Zhang,
691 M., Perlitz, Y., Hernandez-Orallo, J., Li, R., Shabani, A.,
692 Juefei-Xu, F., Dhingra, S., Zohar, O., Nguyen, M. C.,
693 Pondaven, A., Yilmaz, A., Zhao, X., Jin, C., Jiang, M.,
694 Todoran, S., Han, X., Kreuer, J., Rabern, B., Plassart, A.,
695 Maggetti, M., Yap, L., Geirhos, R., Kean, J., Wang, D.,
696 Mollaei, S., Sun, C., Yin, Y., Wang, S., Li, R., Chang, Y.,
697 Wei, A., Bizeul, A., Wang, X., Arrais, A. O., Mukher-
698 jee, K., Chamorro-Padial, J., Liu, J., Qu, X., Guan, J.,
699 Bouyamoun, A., Wu, S., Plomecka, M., Chen, J., Tang,
700 M., Deng, J., Subramanian, S., Xi, H., Chen, H., Zhang,
701 W., Ren, Y., Tu, H., Kim, S., Chen, Y., Marjanović, S. V.,
702 Ha, J., Luczyna, G., Ma, J. J., Shen, Z., Song, D., Zhang,
703 C. E., Wang, Z., Gendron, G., Xiao, Y., Smucker, L.,
704 Weng, E., Lee, K. H., Ye, Z., Ermon, S., Lopez-Miguel,
705 I. D., Knights, T., Gitter, A., Park, N., Wei, B., Chen,
706 H., Pai, K., Elchanany, A., Lin, H., Siedler, P. D., Fang,
707 J., Mishra, R., Zsolnai-Fehér, K., Jiang, X., Khan, S.,
708 Yuan, J., Jain, R. K., Lin, X., Peterson, M., Wang, Z.,
709 Malusare, A., Tang, M., Gupta, I., Fosin, I., Kang, T.,
710 Dworakowska, B., Matsumoto, K., Zheng, G., Sewuster,
711 G., Villanueva, J. P., Rannev, I., Chernyavsky, I., Chen,
712 J., Banik, D., Racz, B., Dong, W., Wang, J., Bashmal, L.,
713 Gonçalves, D. V., Hu, W., Bar, K., Bohdal, O., Patlan,
714
- 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., Sundarapandiyar, 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> [cs]. 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 Hobhahn, M. Forecasting Frontier Language Model Agent Capabilities, March 2025. URL <http://arxiv.org/abs/2502.15850> [cs]. 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> [cs]. 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> [cs]. arXiv:2311.12022 [cs].

- 715 Robison, K. Meta got caught gaming AI benchmarks, April
716 2025. URL <https://www.theverge.com/meta/645012/meta-llama-4-maverick-benchmarks-gaming>.
- 717
- 718 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].
- 719
- 720 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>.
- 721
- 722 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>.
- 723
- 724 Scale AI. Introducing VISTA: A Rubric-based Visual Task Assessment, 2025. URL https://scale.com/leaderboard/visual_language_understanding.
- 725
- 726 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>.
- 727
- 728 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>.
- 729
- 730 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].
- 731
- 732 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].
- 733
- 734 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].
- 735
- 736 Steinhardt, J. AI Forecasting: One Year In, July 2022. URL <https://bounded-regret.ghost.io/ai-forecasting-one-year-in/>.
- 737
- 738 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].
- 739
- 740 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].
- 741
- 742 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.
- 743
- 744 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.
- 745
- 746 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].
- 747
- 748
- 749
- 750
- 751
- 752
- 753
- 754
- 755
- 756
- 757
- 758
- 759
- 760
- 761
- 762
- 763
- 764
- 765
- 766
- 767
- 768
- 769

- 770 White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B.,
 771 Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Dey, S.,
 772 Shubh-Agrawal, Sandha, S. S., Naidu, S., Hegde, C., Le-
 773 Cun, Y., Goldstein, T., Neiswanger, W., and Goldblum,
 774 M. LiveBench: A Challenging, Contamination-Limited
 775 LLM Benchmark, April 2025. URL <http://arxiv.org/abs/2406.19314>. arXiv:2406.19314 [cs].
 776
- 777 Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R.,
 778 Hua, T. J., Cheng, Z., Shin, D., Lei, F., Liu, Y., Xu,
 779 Y., Zhou, S., Savarese, S., Xiong, C., Zhong, V., and
 780 Yu, T. OSWorld: Benchmarking Multimodal Agents for
 781 Open-Ended Tasks in Real Computer Environments. In
 782 Globersons, A., Mackey, L., Belgrave, D., Fan, A., Pa-
 783 quet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances
 784 in Neural Information Processing Systems 38: Annual
 785 Conference on Neural Information Processing Systems
 786 2024, NeurIPS 2024, Vancouver, BC, Canada, Decem-
 787 ber 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.
 788
- 789 Xu, F. F., Song, Y., Li, B., Tang, Y., Jain, K., Bao,
 790 M., Wang, Z. Z., Zhou, X., Guo, Z., Cao, M., Yang,
 791 M., Lu, H. Y., Martin, A., Su, Z., Maben, L., Mehta,
 792 R., Chi, W., Jang, L., Xie, Y., Zhou, S., and Neubig,
 793 G. TheAgentCompany: Benchmarking LLM Agents
 794 on Consequential Real World Tasks, September 2025.
 795 URL <http://arxiv.org/abs/2412.14161>. arXiv:2412.14161 [cs].
 796
- 797 Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi,
 798 Y. HellaSwag: Can a Machine Really Finish Your
 799 Sentence? In Korhonen, A., Traum, D., and Màrquez, L.
 800 (eds.), *Proc. of ACL*, pp. 4791–4800, Florence, Italy,
 801 2019. Association for Computational Linguistics. doi:
 802 [10.18653/v1/P19-1472](https://doi.org/10.18653/v1/P19-1472). URL <https://aclanthology.org/P19-1472>.
 803
- 804 Zhang, A. K., Perry, N., Dulepet, R., Ji, J., Menders,
 805 C., Lin, J. W., Jones, E., Hussein, G., Liu, S., Jasper,
 806 D., Peetathawatchai, P., Glenn, A., Sivashankar, V.,
 807 Zamoshchin, D., Glikbarg, L., Askaryar, D., Yang, M.,
 808 Zhang, T., Alluri, R., Tran, N., Sangpisit, R., Yiorkadjis,
 809 P., Osele, K., Raghupathi, G., Boneh, D., Ho, D. E., and
 810 Liang, P. Cybench: A Framework for Evaluating
 811 Cybersecurity Capabilities and Risks of Language Models,
 812 April 2025. URL <http://arxiv.org/abs/2408.08926>. arXiv:2408.08926 [cs].
 813
- 814 Zhou, L., Pacchiardi, L., Martínez-Plumed, F., Collins,
 815 K. M., Moros-Daval, Y., Zhang, S., Zhao, Q., Huang, Y.,
 816 Sun, L., Prunty, J. E., Li, Z., Sánchez-García, P., Chen,
 817 K. J., Casares, P. A. M., Zu, J., Burden, J., Mehrbakhsh,
 818 B., Stillwell, D., Cebrian, M., Wang, J., Henderson, P.,
 819

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 $\mu_L \sim \text{Beta}(\text{mean} = \frac{0.96 - L_{\min}}{1 - L_{\min}}, \text{sd} = \frac{0.02}{1 - L_{\min}})$ and $\sigma_L \sim \text{Half-N}(\text{sd} = \frac{0.02}{1 - L_{\min}})$.

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 $k_\sigma \sim \text{Half-N}(\text{sd} = 0.005)$.

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 $\xi_\sigma^{\text{base}} \sim \text{Half-N}(\text{sd} = 0.05)$, where N corresponds to the top- N frontier of scores considered ($N = 3$ in this paper).

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

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 $\alpha_\sigma^{\text{raw}} \sim \text{Half-N}(\text{sd} = 0.5)$.

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

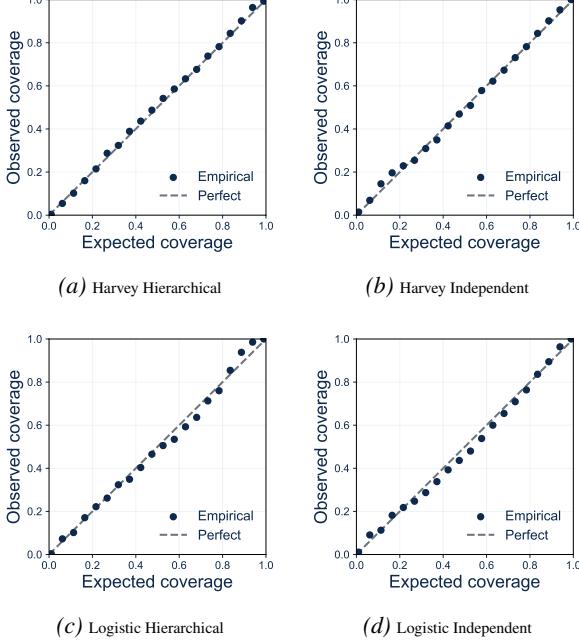


Figure 5. Calibration curves for the four models.

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

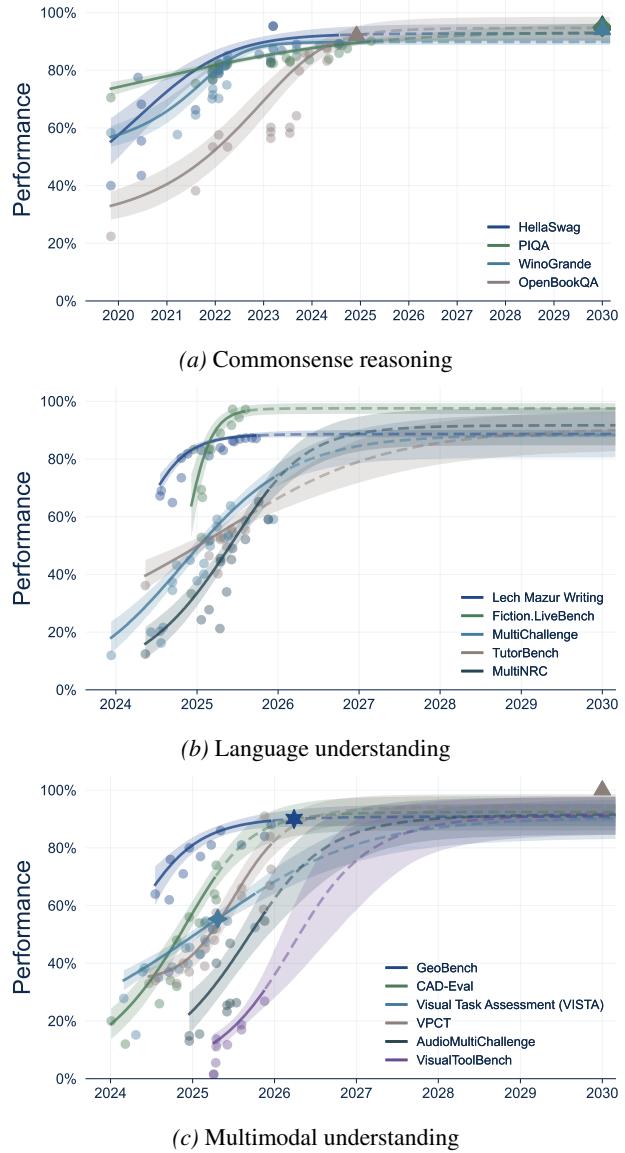


Figure 6. Additional benchmark trajectories. Additional capability categories showing consistent saturation patterns. Commonsense reasoning benchmarks (HellaSwag, PIQA, WinoGrande) are already near saturation; language and multimodal understanding show trajectories consistent with other categories.

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

- 935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
- **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
<i>Domain-Specific Knowledge</i>			
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)
<i>General Reasoning</i>			
SimpleBench	Average Human	83.7%	Philip & Hemang (2024)
<i>Mathematical Reasoning</i>			
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)
<i>Core AGI Progress</i>			
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)
<i>Agentic Computer Use</i>			
OS World	Average Human	72.4%	Xie et al. (2024)
<i>Biology & Chemistry (Dual-Use)</i>			
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)
<i>Commonsense Reasoning</i>			
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)
<i>Multimodal understanding</i>			
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.