

An Agnostic Approach to Statistical Analysis of METR’s AI Time Horizon Data

Jacob Carlson
Harvard University*

October 2026

Abstract

The AI evaluation non-profit METR recently published a series of influential plots that suggested rapid growth in AI agents’ ability to reliably complete long, realistic software and research engineering tasks. To date, most analyses of the data underlying these METR plots, including the original analysis, impose strong (parametric) modeling assumptions to deliver estimates of AI capabilities. In this note, we reanalyze the same underlying data, but this time pursue an “agnostic” statistical analysis of AI capabilities that makes no parametric assumptions about the underlying data generating process (nor very many assumptions at all). Under this reanalysis, *AI progress still looks very fast*, though sampling uncertainty is large enough to make many agents’ capabilities statistically indistinguishable from one another, and various takeaways from the original METR plots appear to be moderated (e.g., we only confidently conclude that Claude Mythos has a 50%-time horizon above 4 hours). Practical takeaways for future evaluations are also discussed.

1 Introduction

The non-profit “Model Evaluation and Threat Research” (METR) recently published a series of AI agent evaluations and associated plots that received a great deal of attention from researchers and policymakers, convincing many that AI progress on difficult, real-world (software and research engineering) tasks was moving exponentially fast in terms of a particular measurement: the *task-completion time horizon* of an AI, defined as “the task

*Email: jacob_carlson@g.harvard.edu. I thank the work of Drew Nguyen and William Fithian for directly inspiring this note. I thank Isaiah Andrews for helpful comments and feedback. I thank OpenAI’s Astra for research assistance, namely checking for errors in and repairing mathematical claims. First draft posted on October 28, 2026.

duration (measured by human expert completion time) at which an AI agent is predicted to succeed with a given level of reliability” [METR, 2026]. Common choices reported by METR are 50% and 80% reliability, inducing the 50%-time horizon and 80%-time horizon plots that METR maintains on their website: <https://metr.org/time-horizons>.¹

The introduction of Nguyen and Fithian [2026] provides a nice, succinct discussion of the social impact of this plot, and, in light of this impact, extra motivation to pursue analysis of the data collected by METR with renewed statistical rigor.

In contrast to the original METR report [Kwa et al., 2026] and methodological extensions of the original report’s analysis developed thereafter [Barry, 2026, Nguyen and Fithian, 2026], this note will be taking a look at METR’s data with a different statistical philosophy: that of (statistical model) “agnostic” methods that aim to impose as few unverifiable structural assumptions about the data as possible, i.e., to remain nonparametric, focus on inference for parameters defined by an unknown distribution, and otherwise let the data speak as much for itself as possible [Angrist and Pischke, 2009, Lin, 2013, Aronow and Miller, 2019].

More specifically, in this note we will make no parametric statistical modeling assumptions to do inference on time horizons and related capabilities-relevant quantities, unlike previous works.² Instead, we will simply let carefully defined estimands, a coherent notion of a sampling experiment in the setting of this data, and ordinary large sample statistical theory guide our statistical inferences. By doing so, our estimands remain highly legible and credible (driven by data and not assumptions), and we can make clear what are considered to be relevant sources of uncertainty underlying the METR data—and to what extent this uncertainty shapes our conclusions about the pace of AI progress in the recent past. Ultimately, this all boils down to doing careful statistical inference on some averages of data that METR collected—no major methodological innovations are required for the present analysis.

Before proceeding, we would like to specifically and explicitly credit Nguyen and Fithian [2026] for inspiring this note, and for offering a very clear statistical characterization of what existing work on the METR data has done. The basic statistical setup and notation in this paper will follow that of Nguyen and Fithian [2026], but will depart from the type of parametric statistical modeling exercise they pursue.

¹Plots viewable online as of October 24, 2026.

²Barry [2026] is the only technical exception, but the nonparametric approaches investigated are toward learning conditional expectation functions, do not consider inference on functionals of the task distribution, and have no guarantees; as such, they live in the same family of approaches as parametric success curve fitting methods.

2 Framework

2.1 Motivation

The basic goal of most existing analyses of the METR data is to do estimation and inference on the $X\%$ -time horizon for some set of AIs using a dataset of recorded AI success or failure at some set of tasks (presented in Figure 1 of [Kwa et al. \[2026\]](#)).

The $X\%$ -time horizon can be defined nonparametrically as the inverse of a monotonic conditional expectation function called the “success (rate) curve” (see Definition 2.1 of [Nguyen and Fithian \[2026\]](#), or a restatement in Section 2.2). However, in all popular analyses (including [Nguyen and Fithian \[2026\]](#)), the focus is on estimates of the $X\%$ -time horizon derived from parameters fit in an Item Response Theory (IRT)-style parametric statistical model of (some part of) the underlying data generating process. The benefit of this approach is that it is conceptually straightforward and allows for point estimation of the time horizons. The drawback is that one needs to believe the parametric statistical model is either correct (which would be heroic) or misspecified yet “sufficiently good” (at recovering the true success curve) in order to believe the resulting estimates for time horizons. However, it is unclear what a satisfying definition of “sufficient” or metric for “good” should be, leading previous analyses to empirically investigate the fit of many different candidate statistical models across a variety of fit scoring metrics, with no very clear winners nor guarantees about approximation of the true success curve (e.g., in [Nguyen and Fithian \[2026\]](#), all newly proposed models have 0 score improvement in their 95% confidence intervals). Such exercises are further complicated by the well-known “Rashomon effect” pathology of fitting parametric models [[Breiman, 2001](#)].

Here, we will avoid the conceptual, statistical, and computational challenges related to searching for “sufficiently good” misspecified parametric models by building a framework for (statistical) model-agnostic, nonparametric estimation and inference on a related capability quantity from the original METR plot paper, a conditional version of the “expected success rate” of AIs for a specific task distribution.³ In and of itself, this quantity will provide meaningful, assumption-lean insights on AI capabilities—as tacitly acknowledged by its variants appearing in previous work (see, e.g., Figures 3 and 12 of [Kwa et al. \[2026\]](#) and Figure 4 of [Nguyen and Fithian \[2026\]](#))—but it will also yield direct hypothesis tests of AI time horizons.⁴ The drawback will be that we do not get precise point estimates for time horizons. (This will be discussed further in Section 2.5.)

In the following framework, we also abstract from concerns about construct validity; see

³In a sense, one could think of existing approaches as doing sound inference on “pseudo-horizons” defined by some parametric statistical model, which have unknown bias with respect to true, nonparametrically defined horizons. In this note, we will focus directly on inference related to the true horizons.

⁴Under the assumption that task success is monotone decreasing in measured human completion time (an assumption made in all previous work).

Nguyen and Fithian [2026] for a nice discussion of this as well. That is, we will assume, by revealed preference of AI commentators, that the measures METR collected are valuable at informing AI capabilities. If one were to believe they were not valuable raw measures, the agnostic analysis described hereafter could still be applied to an alternative dataset with more construct validity.

2.2 Setup

The crux of the agnostic framework is that the METR plot dataset can be construed as the result of sampling tasks iid from *some* task distribution that is informative about AI capabilities, and that we can use this sample to do inference on functionals of that distribution without modeling any aspect of it explicitly. That is, we imagine there is some (very large) set of possible software tasks of interest, and the tasks that appear in the METR dataset are a small random sample from that large, target population of tasks of interest.⁵ Most readings of how METR understands their task dataset supports this framing, which echoes the superpopulation evaluation framing of Miller [2024].⁶

We may denote a software task indexed i as the random object $S_i \in \mathcal{S}$, for some abstract space $\mathcal{S}$ that specifies all relevant task characteristics, and we assume that $S_i \stackrel{iid}{\sim} Q$, where Q is a target distribution of software tasks (implied by the research tastes and interests of the METR researchers who compiled the dataset). We sample n tasks from Q . Then for each task $i \in [n] := \{1, 2, \dots, n\}$, we run AI agents $j \in \mathcal{J}$ on the task and record success or failure for each of the r runs (per some pre-specified criteria), denoted by $Y_{ijr} \in \{0, 1\}$. Per information about how METR ran these assessments, we assume we can write that $Y_{ijr} = g_j(S_i, \xi_{ijr})$ for some measurable function g_j , where ξ_{ijr} is some independent, run-level noise that affects task success, and thus we assume $Y_{ijr} \mid S_i \stackrel{iid}{\sim}$ for each (i, j) pair. This is broadly consistent with the setup and notation of Nguyen and Fithian [2026].

Per METR’s original setup, consider that 8 runs are allocated for each task, so we can also view the sampling experiment as drawing

$$(S_i, \{Y_{ijr}\}_{j \in \mathcal{J}, r \in [8]}) \stackrel{iid}{\sim} P$$

where P is the implied distribution that combines Q and the run-level noise.⁷

⁵Seen this way, the task “family”—which is also recorded by METR—is relevant only insofar as it defines the make-up of the task distribution we are sampling from; there is no sense in which we need to adjust for it from a statistical dependence perspective.

⁶E.g., “Overall we believe it’s likely the new tasks are drawn from a slightly different distribution of difficulty”; “The process of updating our task suite has highlighted the importance of defining the distribution of tasks which we wish to measure performance over” [METR, 2026].

⁷In practice, not all 8 runs get completed for all models and tasks in the original METR analysis, but this is not substantive with respect to the framework discussed, assuming that run failure is independent of evaluation relevant variables.

Specifically, for each fixed model $j \in \mathcal{J}$, we are interested in estimating the nonparametrically founded parameter

$$\theta_j = E[Y_{ijr} \mid b(S_i) = 1] = E[E[Y_{ijr} \mid S_i] \mid b(S_i) = 1]$$

where $b : \mathcal{S} \rightarrow \{0, 1\}$ is some binary-valued function of the full task information, e.g., the task completion time for humans falling in the 4-16 hours bin. The second equality follows by the tower property of conditional expectations. In words, owing to the second expression, we might call θ_j the “expected task success rate” of AI j at a task with characteristics that fall in some bin defined via b . For shorthand, going forward, denote $B_i := b(S_i)$.

Note that θ_j is not the $X\%$ -time horizon quantity of original interest, which [Nguyen and Fithian \[2026\]](#) define as $p_j^{-1}(X\%)$ where $p_j(t) = E[Y_{ij1} \mid \text{human completion time} = t]$. However, θ_j is both (i) a legible, capabilities-relevant quantity, echoed by the above-mentioned places in existing analyses where a version of it appears and (ii) a key ingredient for further agnostic inferences about the time horizon, e.g., if b defines the 4-16 hours bin, and we reject based on METR’s data that $\theta_j < 0.5$, then we also reject that the 50%-time horizon for model j is below 4 hours under a success curve monotonicity assumption. It’s worth noting here that there are very likely to be agnostic approaches to point estimation of time horizons—and associated inference results—under a monotonicity assumption, which could be the subject of future work (this will be discussed again briefly in [Section 2.5](#)). However, for the sake of developing a conceptually straightforward, baseline agnostic approach, we will proceed in this note by working with simple averages and doing appropriate inference on their levels and comparison.

2.3 Estimation

Further notice that, defining $Z_{ij} := \frac{1}{8} \sum_{r=1}^8 Y_{ijr}$ then just by linearity of (conditional) expectations

$$E[Z_{ij} \mid S_i] = E[Y_{ijr} \mid S_i]$$

almost surely, and so again using the tower property

$$\theta_j = E[E[Y_{ijr} \mid S_i] \mid b(S_i) = 1] = E[E[Z_{ij} \mid S_i] \mid B_i = 1] = E[Z_{ij} \mid B_i = 1] = \pi_b^{-1} E[Z_{ij} B_i]$$

where $\pi_b := E[B_i]$. The natural plug-in estimator is then

$$\hat{\theta}_j := \hat{\pi}_b^{-1} \cdot \frac{1}{n} \sum_{i=1}^n Z_{ij} B_i$$

where $\hat{\pi}_b := n^{-1} \sum_i B_i$.

2.4 Inference

The following inference results are all based on standard arguments, so formal proofs are omitted for brevity, and the exposition will focus on intuition.

2.4.1 Large Sample Theory

Inference now proceeds by observing that the run-averaged data Z_{ij} are really just iid at the task level i . Thus, the effective sample size is $\pi_b n$, the number of tasks in the bin, but the run count (here 8 for simplicity, but can be arbitrary) will affect our estimator’s overall variance, which will have a component due to task heterogeneity and due to run-level noise.

Specifically, we may group all of our $|\mathcal{J}|$ parameters of interest as the vector $\theta \in [0, 1]^{|\mathcal{J}|}$ and the corresponding estimators as $\hat{\theta}$. By a standard WLLN for iid data (so $\hat{\pi}_b \xrightarrow{p} \pi_b$) and a multivariate CLT for iid data

$$\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow N(0, \pi_b^{-1} \Sigma),$$

where $\rightsquigarrow$ denotes convergence in distribution and, for $Y_{ir} := (Y_{i1r}, \dots, Y_{i|J|r})^T$ and $Z_i := (Z_{i1}, \dots, Z_{i|J|})^T$,

$$\Sigma = \text{Var}[Z_i \mid B_i = 1] = \underbrace{\text{Var}[E[Y_{ir} \mid S_i] \mid B_i = 1]}_{(I)} + \underbrace{E[\text{Var}[Z_i \mid S_i] \mid B_i = 1]}_{(II)}.$$

The second equality follows by the law of total variance, and decomposes the asymptotic variance into a component driven by task success heterogeneity (I) and within-task noise (II). Adding runs to the dataset therefore only affects the term (II), and one can use this decomposition as a basis for weighing adding more runs versus more tasks to a future dataset.

Note that this is equivalent to thinking of inference on functionals of the run-level outcome distribution with task-level clustering (in this setup with an equal number of runs per task). Regardless of framing, statistical dependence in Y_{ijr} across models and within tasks is left *unrestricted*—the data informs it directly, and we need make no additional assumptions.

That this is standard large sample (asymptotic) theory suggests that we need a reasonably large number of tasks to fall in the bin defined by b .⁸ This in turn suggests that organizations interested in AI evaluations may want to focus on collecting large amounts of tasks for bins of greatest interest.

2.4.2 Simultaneous Confidence Sets and Multiple Testing

Ultimately, we want to be able to easily infer significant changes across multiple $\hat{\theta}_j$ s, or run hypothesis tests on them all in a way that accounts for multiple testing ($|\mathcal{J}|$ is not tiny, 10-30 in the METR plot case). We can handle both by appeal to constructing a joint or

⁸In the 4-16 hours bin, there are 46 tasks.

“simultaneous” confidence region for the elements of θ (see for reference, e.g., [Romano and Wolf \[2007\]](#), [Chernozhukov et al. \[2013\]](#), [Montiel Olea and Plagborg-Møller \[2019\]](#)). This will be the inversion of a multiple testing procedure that controls the familywise error rate (FWER)—the probability of making any false rejections were we to run a hypothesis test based on every element of θ .

Specifically, a large sample valid simultaneous confidence interval has the guarantee that

$$\liminf_{n \rightarrow \infty} P(\mathcal{C}_n \ni \theta) \geq 1 - \alpha$$

for some chosen α , where $\mathcal{C}_n$ is a data-dependent set (it is random, whereas θ is fixed). A straightforward way of constructing such a set of intervals is with a max-t style method, the intuition for which is as follows. If we wanted to test the null $H_{0,j} : \theta_j = 0.5$ for all $j \in \mathcal{J}$ using a common cutoff value c , rejecting $H_{0,j}$ when test statistic $|T_j| > c$ (where $T_j := \sqrt{n}(\hat{\theta}_j - 0.5)$), then for the set of indices of hypotheses with true nulls $I(P)$,

$$P(\text{any false rejections}) = P(\max_{j \in I(P)} |T_j| > c) = P(\max_{j \in I(P)} \sqrt{n}|\hat{\theta}_j - \theta_j| > c) \leq P(\max_{j \in \mathcal{J}} \sqrt{n}|\hat{\theta}_j - \theta_j| > c).$$

So if we set c to be the $1 - \alpha$ quantile of $\max_{j \in \mathcal{J}} |\sqrt{n}(\hat{\theta}_j - \theta_j)|$ (or its feasible approximation with a valid bootstrap), then we obtain “strong” FWER control (in large samples, when we approximate it).⁹

Moreover, by the inversion/duality of confidence intervals and hypothesis tests, we may consider the set

$$\{\vartheta : \max_{j \in \mathcal{J}} |\sqrt{n}(\hat{\theta}_j - \vartheta_j)| \leq c\} = \bigcap_{j \in \mathcal{J}} \{\vartheta : |\sqrt{n}(\hat{\theta}_j - \vartheta_j)| \leq c\} = \prod_{j \in \mathcal{J}} \left[\hat{\theta}_j - \frac{c}{\sqrt{n}}, \hat{\theta}_j + \frac{c}{\sqrt{n}} \right] := \mathcal{C}_n$$

to be a $1 - \alpha$ simultaneous confidence set for θ . That is, the frequentist probability that all of the θ_j are jointly contained by these intervals is $1 - \alpha$ in large samples, under mild consistency assumptions for our estimated c . To compute this simultaneous confidence set in practice in large samples, a (multiplier) bootstrap for the $1 - \alpha$ quantile of $\max_{j \in \mathcal{J}} |\sqrt{n}(\hat{\theta}_j - \theta_j)|$ will suffice at the (iid) task-level. That is, in the given sampling thought experiment, there is no theoretical motivation to do resampling at any other level (more on this in [Section 2.5](#)). In practice, we would also want to studentize these intervals (this is the true max-t method).

By constructing such a simultaneous confidence set, then, with high probability in a frequentist sense, all θ_j are covered by the random intervals. This is a particularly “plot-friendly” guarantee, because, in the high probability “good event,” all intervals cover, so we may simply inspect if the confidence intervals of any two $\theta_j, \theta_{j'}$ do not overlap to conclude there is a statistically significant difference between them. Overlap, on the other hand,

⁹“Strong” in the sense that this holds for any possible $I(P)$.

means that we cannot reject based on these CIs that $\theta_j = \theta_{j'}$.¹⁰ Just by inspection, we can also infer which parameters θ_j we can reject are above or below, e.g., 0.5, which are the parameters whose intervals lie completely above or below 0.5, respectively, which also comes with the high probability frequentist guarantee that there aren’t any false rejections (FWER control). Furthermore, these simultaneous intervals guard against post-selection inference problems, which is to say looking at the data to select some subset of indices $\hat{\mathcal{J}}$ to report point estimates and CIs for; again, on the “good event,” all intervals cover, and so any (even adversarial) choice of indices has correct simultaneous coverage.

These properties of simultaneous coverage are important to have in CIs insofar as joint statistical guarantees over time/agents are key to forming scientific insight about the rate of AI progress—especially when a very specific hypothesis test across a specific small set of models is not reasonably specifiable ex-ante. This is to say that when we look at a plot of these estimates with simultaneous confidence sets, the sort of things we would intuitively like to be read directly off of the plot (e.g., non-overlapping intervals means significant difference, intervals not containing a point have a correction for looking at many intervals) are actually supported by the underlying statistical theory.

2.5 Discussion

One of the reasons why this analysis gets away with making very few assumptions is that it does not attempt to estimate, invert, or otherwise do inference on a conditional expectation function of a continuous variable—the success curve. An implicit argument of the present analysis, then, is that we need not tackle this more difficult estimation challenge in order to still learn something relevant to AI capabilities and time horizons from the METR data.

Put another way, analyzing discrete time bins of AI success provides a meaningful pathway to model-agnostic analysis that informs time horizons; learning expected task success for every possible human completion time on a segment of real line is both difficult and possibly overkill. Given how much uncertainty already lives in a fairly coarse time bin, one might be skeptical of inferences at specific points (infinitesimally small bins) in the space of human completion time that appear quite certain under statistical modeling assumptions. That said, methods related to (adaptive) shaped constrained inference for regression functions or (adaptive, conditional) moment inequalities may also prove useful in surfacing a more agnostic estimation and inference strategy that directly yields point estimates of time horizon quantity (c.f., [Andrews and Shi \[2013\]](#), [Armstrong \[2015\]](#), [Chetverikov \[2018\]](#), [Chetverikov et al. \[2018\]](#)). Future work might consider this.

¹⁰Simultaneous tests of pairwise differences could improve power to separate model capabilities, by leveraging estimated covariances that can’t be read off of the plot directly, under appropriate multiple testing corrections. As such, failing to reject based on overlap in the simultaneous $\hat{\theta}_j$ plot is conservative. More on this in [Section 3](#).

If one’s goal is to learn *latent* parameters attached to some statistical model and do estimation and inference on those parameters, as using item-response theory (IRT) style methods might suggest, then parametrically modeling success curves and assessing fit may be the most natural approach. However, the agnostic philosophy would be quite skeptical of both the interpretability and intrinsic meaningfulness of such latent parameters, which could have less legible construct validity than imperfect but highly interpretable task success measures.

It’s also worth noting that quantified uncertainty seems to be a meaningful part of how the original METR author(s) and subsequent researchers interpret the METR plot results [Kwa, 2026, Barry, 2026]. However, the original analysis of the data in Kwa et al. [2026] provides no formal justification for their hierarchical bootstrapping procedure, which drives uncertainty quantification in the headline plots. The agnostic analysis, by virtue of being a question about parameter inference, forces clarity and justification for formal confidence statements.

Lastly, the agnostic analysis is trivial to produce from a computational perspective, as opposed to more complicated parametric model fitting approaches (c.f., discussion of Model 2 of Nguyen and Fithian [2026], or needing to bootstrap many, many logistic regressions in Kwa [2026]).

3 Empirics

3.1 Analysis of the 4-16 Hour Human Completion Time Bin

We now implement the framework discussed in the previous section, using METR’s publicly available v1.1 data [METR, 2026].

Specifically we set b to pick out tasks with human completion times in the $(4, 16]$ hour interval (the smallest coherent frontier bin with reasonable amounts of data). We focus on agents from the “reasoning era” onward, which is an ex-ante choice of agents that would be competitive in this time bucket. There are 46 tasks on which all reasoning agents complete their assessments.

Figure 1 plots the conventional, large sample valid, parameter-by-parameter Wald intervals as well as the discussed simultaneous intervals and parameter point estimates; implied FWER control is at $\alpha = 5\%$, and studentization is performed in a multiplier bootstrap.¹¹ A dotted horizontal line indicates an expected task success rate of 50%. Figure 2 plots the same information, but scattered by release date of model.

The most salient insights from these plots seem to be:

¹¹In practice, we could also estimate log-odds (valid via a conventional delta method argument in large samples) and then apply an expit transformation to produce intervals that always remain within the support of the data.

1. The frontier AI progress time trend (as defined by agents that led performance at any point in time) appears to be linear with a slope of approximately 4 percentage points increase per month in expected 4-16 hour task success. Completely speculatively, if this trend continued, it would saturate the “4-16 hour benchmark” by roughly the start of 2027.
2. Only Claude Mythos Preview has a simultaneous interval sitting entirely above 0.5, indicating we would reject its time horizon is below 4 hours under a monotonicity assumption—we fail to reject this hypothesis (simultaneously) for any other model since the reasoning era started. Similarly, using monotonicity, an interval entirely below 0.5 bounds the time horizon from above by 16 hours, which is true of models released before GPT-5.
3. Using these simultaneous intervals for $\hat{\theta}_j$ s, one cannot statistically distinguish the expected task success rate of Mythos from any model post Claude 4.1 Opus, as indicated by overlapping simultaneous confidence intervals. If we instead apply the agnostic framework to doing simultaneous inference on $\theta_j - \theta_{j'}$ for all unordered pairs $j, j' \in \mathcal{J}$ (105 comparisons in total), inference on changes in agent success is more powerful, but we still can’t distinguish Claude Mythos from Claude Opus 4.6—see Figure 4 in Section 5.1.
4. In general, most AI agents released in a 6 month window or so of one another cannot be teased apart with the existing METR data, again based on insights directly communicated by the simultaneous confidence intervals in Figures 1 or 2, as well as the more powerful simultaneous pairwise difference inferences plotted in Figure 4 in Section 5.1 (similarly true for the rank inference plot derived from it, Figure 5, in the spirit of Mogstad et al. [2024]). Evidently, more tasks would be needed to separate models 6 months or so apart.

3.2 Time Horizon Plot Redux

One way of nonparametrically producing a METR-style time horizon plot directly using our framework from Section 2 is to expand simultaneous inference to all agents in all human time completion bins (that is, bins growing by approximately $4\times$ human time completion), and then plot inferred frontier upper and lower bounds for time horizons over calendar time (again, under a monotonicity assumption). This yields the plot in Figure 3.

As can be seen, Figure 3 is much coarser in its information content than the original Figure 1 of Kwa et al. [2026], reflecting sampling variability in the underlying dataset and the fact that we do not point identify or point estimate time horizons. It is consistent with more powerful inference on the 4 hour time horizon lower bound per Figures 1 or 2, and

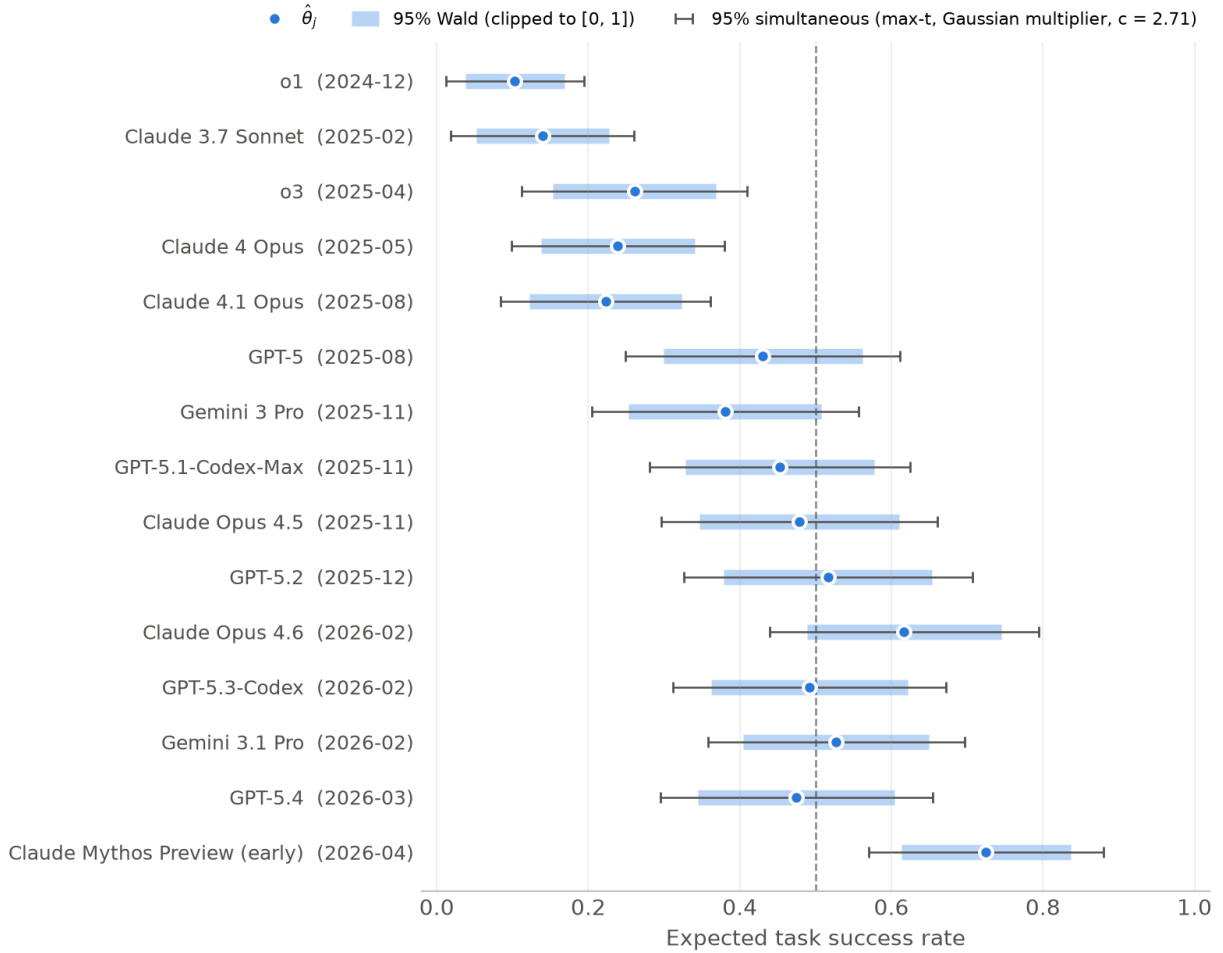


Figure 1: Expected task success rate by AI agents in the post-reasoning era, based on v1.1 of METR’s time horizon dataset.

reveals that Mythos’s release is meaningful in that it was the first model to push the 50%-time horizon lower bound away from 15 minutes since the reasoning era started—though the 80%-time horizon lower bound is meaningfully pushed already by GPT-5 (and the upper bound pushed by Claude Opus 4.5 and then again by Claude Opus 4.6). Moreover, this plot communicates that we cannot rule out (using the METR v1.1. data and this simple agnostic approach) that the 50%-time horizon has been between 4 and 16 hours for the *entire* reasoning era.

Recreation of the entire time horizon plot costs power as the price of simultaneous inference (multiple testing) for all agents in all human time completion bins. As such, conditional on living in a world with agent capability priors informed by the original METR plots, more powerful benchmark-style analyses that still speak to time horizons, such as those in Figures 1 and 2, may prove more attractive going forward.

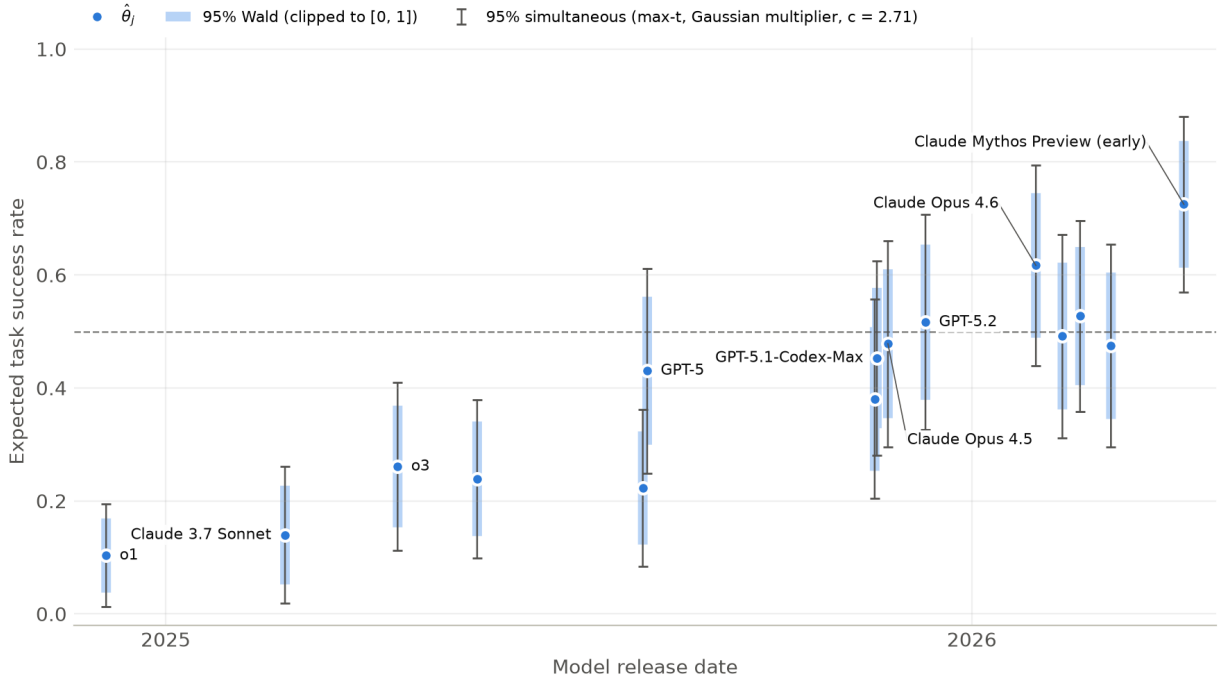


Figure 2: Expected task success rate by AI agents in the post-reasoning era, scattered with release date, based on v1.1 of METR’s time horizon dataset.

4 Takeaways

This reanalysis aims to see just how much we can learn about AI progress from the dataset that produced the famous METR plots using “agnostic” methods that make as few assumptions as possible. The headline takeaways are:

1. Methodological Takeaways: By treating the METR data (or data like it) as inputs to a model-agnostic inference problem, we still unlock rich insight about AI progress and time horizons, but retain a high degree of legibility and credibility in the results: the meaning of the object of interest being estimated, the population that it is relevant with respect to, and rigorous uncertainty quantification all become clear and available under few assumptions. There is no need to search for the next best statistical model of the data. We lose point estimates for time horizons but still have assumption-lean bounds, and there is further reason to believe future work could develop more agnostic point estimates as well.
2. Empirical Takeaways: **AI progress still looks very fast**, even with an agnostic approach. That said, some takeaways from the original METR plot(s) seem moderated or qualified. Namely, using METR’s v1.1 data, Claude Mythos appears to be a somewhat less pathbreaking point of progress (especially relative to Claude Opus

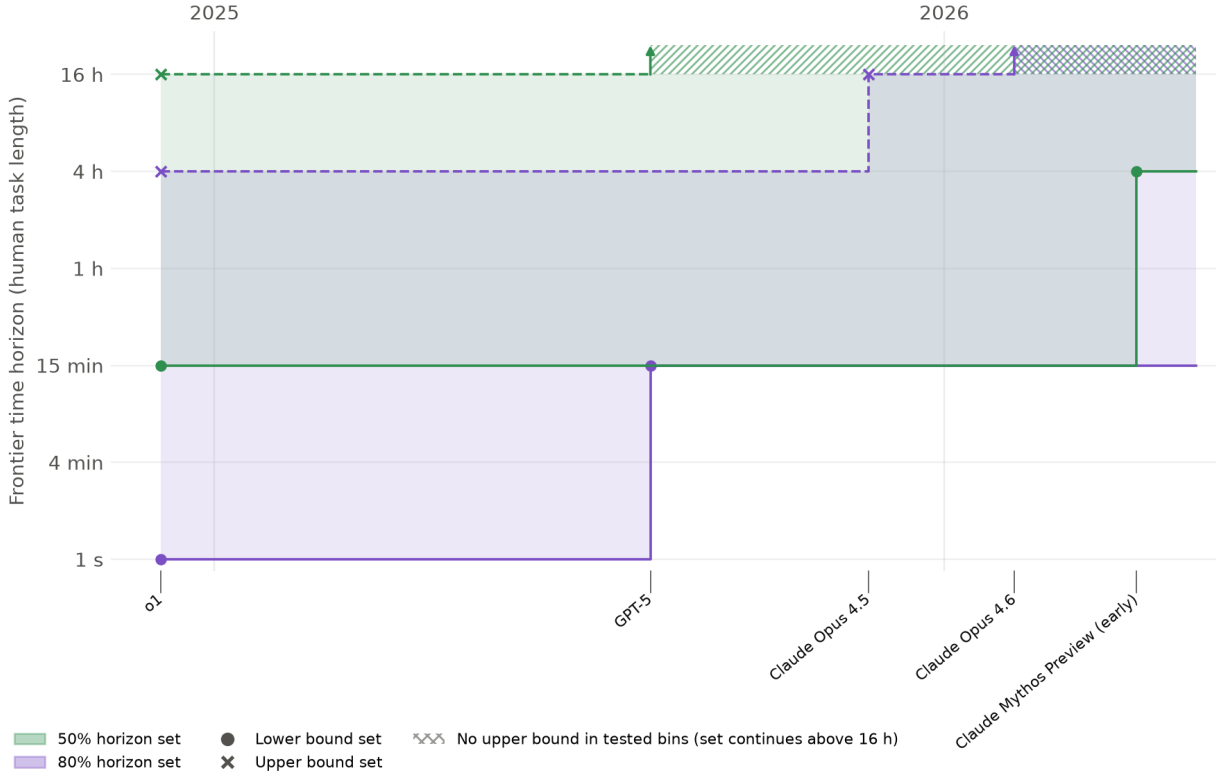


Figure 3: Nonparametrically estimated frontier time horizons for AI agents in the post-reasoning era, based on v1.1 of METR’s time horizon dataset and using 95% simultaneous intervals for expected task success rate in all time bins. In increasing human completion times, bins contain 82, 38, 36, 22, and 46 tasks, respectively.

4.6); agent performance is rigorously shown to be statistically indistinguishable in multi-month time ranges; and frontier 50%-time horizon inferences only speak to lower bounds recently rising above 4 hours (and 80%-time horizon lower bounds are still at 15 minutes as of Mythos Preview).

3. Practical Takeaways: The results in this note also directly suggest how to think about adding more runs to the benchmark versus adding more tasks in order to reduce uncertainty (owing to the variance decomposition) and how to collect tasks of certain lengths going forward to inform rigorous, more granular time horizon inferences.

5 Appendix

5.1 Simultaneous Inference on Pairwise Differences and Ranks

The simultaneous intervals in Figures 1 and 2 yield conservative inference on differences in θ_j s. To improve power, while still accounting for multiple testing, we can directly apply the framework of Section 2 to the vector of all unordered paired differences in expected task success rate, that is the vector collecting $\beta_{j,j'} := \theta_j - \theta_{j'}$ for all unordered pairs $j, j' \in \mathcal{J}$. We can visualize the results of doing so in Figure 4 (read by looking at the y-axis grid labels and working down), in which we see that we fail to reject Claude Opus 4.6’s expected task success rate is different from Claude Mythos’s, and further notice that—as Figure 1 suggested—most models before Mythos cannot be distinguished from one another in 4-6 or so month ranges.

Per Mogstad et al. [2024], we can use this sort of pairwise difference information with simultaneous coverage to directly create a simultaneous confidence region for expected task success rate *rank* of agents, depicted in Figure 5. This plot makes it even clearer that differentiating agent performance is difficult in most of the recent past using the METR data, and also suggests 3 reasoning eras that we have smoothly transitioned between: the first spanning OpenAI’s o1-o3, the next spanning Gemini 3 Pro through Claude Opus 4.6, and another then beginning with Claude Mythos Preview.

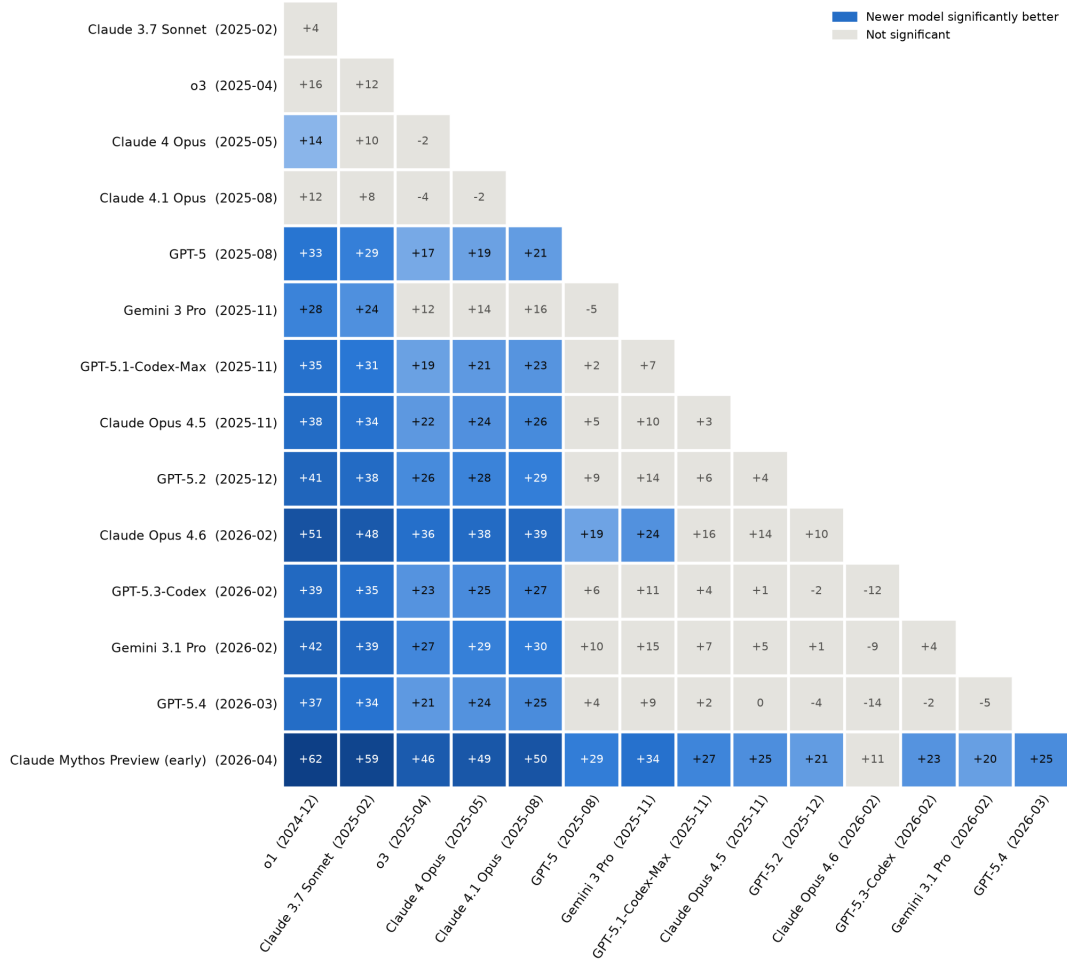


Figure 4: Differences in expected task success rate of AI agents in the post-reasoning era, based on v1.1 of METR’s time horizon dataset, FWER control at 95% via max-t.

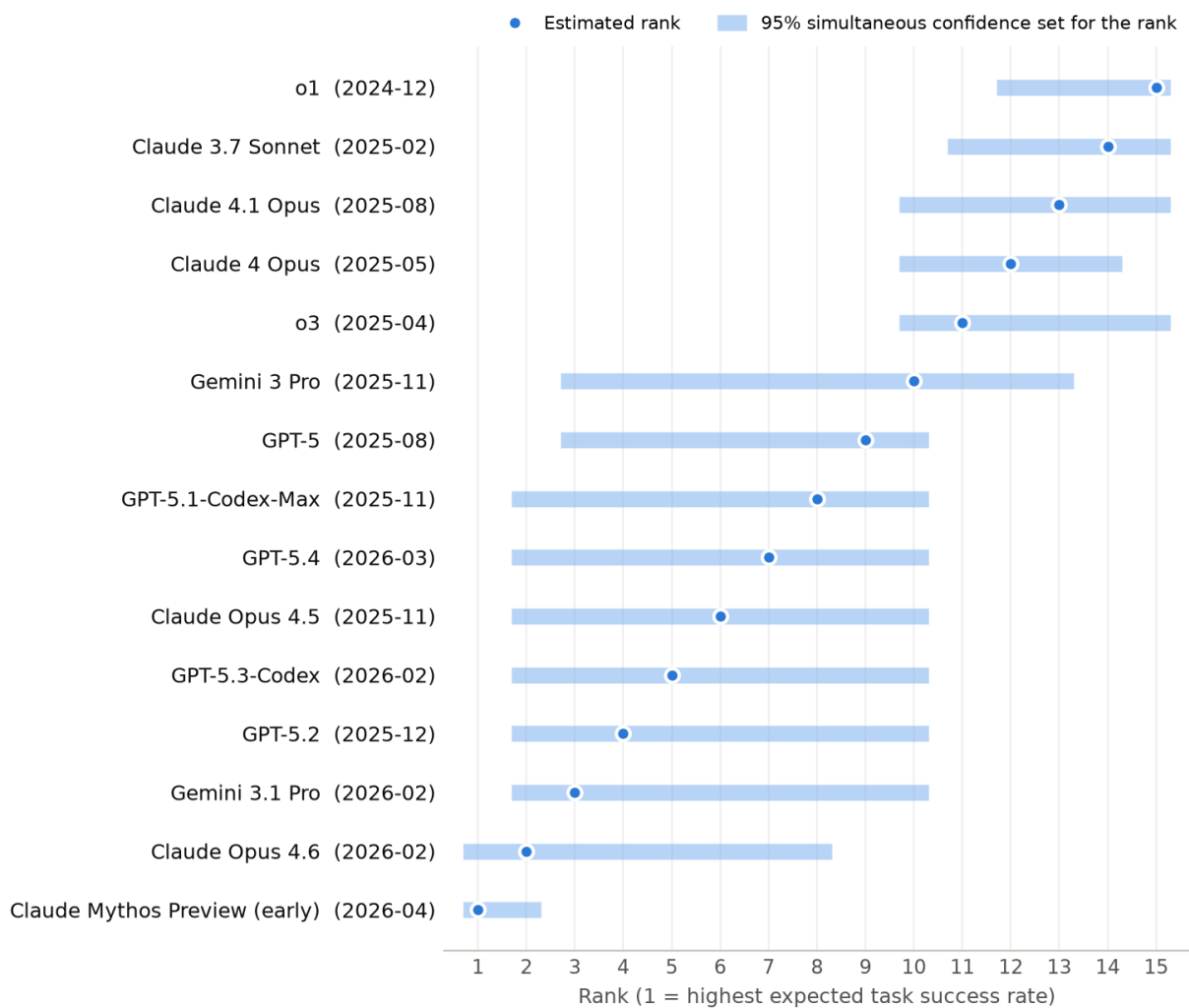


Figure 5: Ranks of expected task success rate by AI agents in the post-reasoning era, in the style of Mogstad et al. [2024], based on v1.1 of METR’s time horizon dataset.

References

- METR. Task-completion time horizons of frontier ai models. <https://metr.org/time-horizons/>, 09 2026.
- Drew T. Nguyen and William Fithian. On the estimation and validity of AI time horizons—a statistical look at the METR plot, October 2026. URL <https://arxiv.org/abs/2610.12466>. Version Number: 1.
- Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Chris Painter, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI Ability to Complete Long Software Tasks, July 2026. URL <https://arxiv.org/abs/2503.14499>. Version Number: 4.
- Alexander Barry. Impact of modeling assumptions on time horizon results. <https://metr.org/notes/2026-03-20-impact-of-modelling-assumptions-on-time-horizon-results/>, 03 2026.
- Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press, Princeton, NJ, 2009. ISBN 9780691120355.
- Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman’s critique. *The Annals of Applied Statistics*, 7(1):295–318, March 2013. doi: 10.1214/12-AOAS583.
- Peter M. Aronow and Benjamin T. Miller. *Foundations of Agnostic Statistics*. Cambridge University Press, 1 edition, January 2019. ISBN 978-1-316-83176-2 978-1-107-17891-5 978-1-316-63114-0. doi: 10.1017/9781316831762. URL <https://www.cambridge.org/core/product/identifier/9781316831762/type/book>.
- Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, August 2001. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1009213726. URL <https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling--The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.full>.
- Evan Miller. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations, 2024. URL <https://arxiv.org/abs/2411.00640>. Version Number: 1.

- Joseph P. Romano and Michael Wolf. Control of generalized error rates in multiple testing. *The Annals of Statistics*, 35(4), August 2007. ISSN 0090-5364. doi: 10.1214/009053606000001622. URL <https://projecteuclid.org/journals/annals-of-statistics/volume-35/issue-4/Control-of-generalized-error-rates-in-multiple-testing/10.1214/009053606000001622.full>.
- Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013. doi: 10.1214/13-AOS1161.
- José Luis Montiel Olea and Mikkel Plagborg-Møller. Simultaneous confidence bands: Theory, implementation, and an application to SVARs. *Journal of Applied Econometrics*, 34(1): 1–17, 2019. doi: 10.1002/jae.2656.
- Donald W. K. Andrews and Xiaoxia Shi. Inference based on conditional moment inequalities. *Econometrica*, 81(2):609–666, 2013.
- Timothy Armstrong. Adaptive testing on a regression function at a point. *The Annals of Statistics*, 43(5):2086–2101, October 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/15-AOS1342. URL <https://projecteuclid.org/journals/annals-of-statistics/volume-43/issue-5/Adaptive-testing-on-a-regression-function-at-a-point/10.1214/15-AOS1342.full>.
- Denis Chetverikov. Adaptive tests of conditional moment inequalities. *Econometric Theory*, 34(1):186–227, 2018.
- Denis Chetverikov, Andres Santos, and Azeem M. Shaikh. The Econometrics of Shape Restrictions. *Annual Review of Economics*, 10(1):31–63, August 2018. ISSN 1941-1383, 1941-1391. doi: 10.1146/annurev-economics-080217-053417. URL <https://www.annualreviews.org/doi/10.1146/annurev-economics-080217-053417>.
- Thomas Kwa. Clarifying limitations of time horizon. <https://metr.org/notes/2026-01-22-time-horizon-limitations/>, 01 2026.
- Magne Mogstad, Joseph P Romano, Azeem M Shaikh, and Daniel Wilhelm. Inference for Ranks with Applications to Mobility across Neighbourhoods and Academic Achievement across Countries. *Review of Economic Studies*, 91(1):476–518, January 2024. ISSN 0034-6527, 1467-937X. doi: 10.1093/restud/rdad006. URL <https://academic.oup.com/restud/article/91/1/476/7008541>.