THE DECLINING HAZARD REPORTED FOR AI AGENTS IS A POOLING ARTEFACT

Anonymous authors
Paper under double-blind review

ABSTRACT

Published fits to METR's time-horizon data put the Weibull shape k below one, so the failure rate falls as tasks get longer, and read it as agents getting better as a task goes on. That decline is an artefact of pooling unequally difficult tasks: a mixed batch of light bulbs shows a falling burnout rate while each wears out. Modelling task difficulty reverses the sign: the shape moves from 0.63 to 1.85, and 17 of 20 models exclude $k \leq 1$, a count no null simulation reached. Across admissible ways of pooling the shape spans 1.56 to 1.85; the data reject the only one consistent with constant hazard. The sign holds on every clock; the magnitude does not: the same runs clocked by tokens, rated task length or elapsed time have median shapes 1.700, 2.045 and 2.625; the ordering repeats on a benchmark sharing no tasks, agents or harness. Humans on the same tasks and clock are indistinguishable from constant hazard; agents exceed them by $\Delta k = 1.225$ $[0.825, 1.725]$. A blind reimplementation from the specification reproduced both headline values. METR's pooled horizons stand. At fixed task difficulty an agent's 80%-reliability length is about half its 50% length; the published shape implies a sixth. Report the sign; report the magnitude as a range across clocks, or not at all.

1 INTRODUCTION

Agentic benchmarks have moved from a single score to a curve. METR's time-horizon evaluation (Kwa et al., 2025) is the clearest instance and the one we analyse: every task in the suite is rated by the time a human expert needs to finish it, agents are run across the suite, and each model's success rate is fitted as a function of that rated time.

The number usually quoted is a horizon, the task length at which a model succeeds half the time. The curve's *shape* carries a second claim: how quickly success falls away as tasks lengthen. A shape parameter below one says the curve's hazard, the share of remaining success lost per unit of length, declines as tasks get longer, though success itself still falls. Published fits to this dataset put the shape below one and read it as a fact about agents: models "don't accumulate errors and get worse; in fact they get better over time" (Hamilton, 2026), and hazard rates for AI agents "decline as a task goes on" (Ord, 2026). That reading has the sign backwards. With task difficulty modelled, agents wear out as a run goes on; humans on the same tasks and clock show no detectable wear.

The sign matters for reliability. METR's published 50% and 80% horizons are read off the pooled curve, and nothing here moves them. The shape sets their conditional counterpart: the ratio of the length an agent manages four times in five to the length it manages half the time, a function of k alone. At $k = 0.63$ it is 0.165, about a sixth; across the corrected range $k = 1.56$ to 1.85 it is 0.484 to 0.542, about half. The published shape understates it by 2.9–3.3$\times$ ($[2.86, 3.71]$ at $k = 1.85$), a reading at fixed difficulty that needs the length-invariance assumption of §4.

The published sign comes from a choice made before any model is evaluated and almost never reported: what to pool over. Agentic benchmarks fit a single curve across a suite whose tasks differ enormously in how hard they are. Pooling items of unequal difficulty produces a population hazard that can fall even while the hazard of every item in the pool rises, because the fragile items fail early and whatever is still running is disproportionately robust. The effect is classical: a mixed batch of light bulbs can show a declining burnout rate though every bulb in it is wearing out (Gurland & Sethuraman, 1995; Vaupel et al., 1979). A benchmark suite holding both near-impossible and routine tasks is such a batch.

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

A second choice, made just as quietly, sets the magnitude. The length of a task is not one quantity. The same run can be indexed by the human time its task was rated at, by the wall-clock seconds the agent spent on it, or by the tokens the agent generated. These are three different questions about one run, and the literature treats them as interchangeable wherever it distinguishes them at all.

(RQ1) Does the published declining-hazard result survive an estimator that models task difficulty, and if it does not, can data of this design identify the corrected sign? **(RQ2)** How far does the fitted shape move with the clock a run is indexed by, and how does that movement compare with the pooling correction? **(RQ3)** Is rising within-run hazard specific to agents or to the measurement, given a human control that worked the same tasks on the same clock? The objective is to separate the part of a benchmark curve's shape that these data identify from the part that belongs to an analyst's undocumented choice, so the second is not mistaken for a finding. A single replacement value of k is not what this paper supplies.

We answer all three on METR's data (65,637 runs, 52 models over two releases, 20 in the second), with two further benchmarks as external checks. The estimator is conventional: the same Weibull family the field fits, with a mixing distribution over tasks estimated alongside it. That mixture can be fitted here because agentic benchmarks run the same task repeatedly against the same model; single-run evaluation cannot separate a hard task from an unlucky one. The estimator was frozen before application and given no per-dataset tuning.

Priority. Ord (2025) raised the confound, and item-response models of task difficulty already exist for these benchmarks (§2); neither tested the sign. Ours is that test, its validation against a measured heterogeneity distribution, the human control and the clock result.

CONTRIBUTIONS

- **The sign of the published shape reverses once task difficulty is modelled**, from 0.63 to 1.85, and we state what this design supports: the sign, not the point. 17 of 20 models exclude $k \leq 1$ against a joint null that never reached that count, and the corrected shape spans 1.56 to 1.85 across admissible aggregations, the one specification that does not exclude constant hazard being rejected by the data (§6.1).
- **The clock moves the magnitude by about as much as the correction does.** The same runs give 1.700, 2.045 and 2.625 on tokens, rated task length and elapsed time; the ordering recurs on a benchmark sharing nothing with METR, and three candidate mechanisms fail (§6.2).
- **A human control on the same tasks and clock** is indistinguishable from constant hazard; the agent arm exceeds it by $\Delta k = 1.225$ $[0.825, 1.725]$ (§6.3).
- **A blind reimplementation from the written specification alone** reproduced both headline values and surfaced the aggregation choice above, and a support-grid convention our text left open that moves the estimate by 0.23 (§6.6); a rejected-hypotheses table records which check killed each candidate explanation, including three of our own.

The clock gap has no mechanism, and it does not need one to bind: a magnitude that moves by about as much as the pooling correction with a clock choice the literature treats as interchangeable cannot be reported from a single clock, and the gap replicates on a benchmark sharing nothing with the first. The sign of a benchmark curve's shape is stable across every admissible specification and clock tested here; its magnitude is a joint property of the data and that choice, and belongs in a paper as a range across clocks or not at all. Every within-run magnitude here is stated in that form.

2 RELATED WORK

The curve has been fitted before, the confound has been named before, and the statistics is standard. Nobody had tested the confound, and the test reverses the published sign.

2.1 FITS TO THE SAME CURVES

METR's time-horizon report (Kwa et al., 2025) established the object we analyse, per-model success as a function of a task's rated human completion time. Hamilton (2026) fits Weibull and log-logistic

families to it and reports declining hazard; Ord (2026) reports and interprets that finding. Moss (2026) studies the curve's location and task-difficulty spread, Xing et al. (2026) separates level from shape, tests ceiling effects and avoids harness confounding, and Liu et al. (2026) ports the framework to further benchmarks. All three model task difficulty with item-response models, but on the success-probability scale, where hazard shape is not a quantity any of them reports; and none of this work varies the clock (App. A).

2.2 The confound was named before it was fitted

Ord (2023) states the mechanism we quantify, and Ord (2025) raises it against METR's data, eight months before the declining-hazard fits it anticipates. The concern was therefore on the record and unresolved in either direction, because only a fitted mixture resolves it.

A Weibull whose hazard is scaled by a unit-variance gamma draw *is* the log-logistic, so the slight goodness-of-fit edge reported for the log-logistic over the Weibull (Ord, 2026; Hamilton, 2026) is the direction a heterogeneous population predicts. It was read as a statement about which curve to draw; only a fitted mixture says whether it is also one about what the population contains.

2.3 Where the statistics comes from

We claim none of the machinery. Mixtures of increasing-hazard distributions can have decreasing population hazard (Gurland & Sethuraman, 1995); frailty, an unobserved nonnegative multiplier on an item's hazard, formalises this (Vaupel et al., 1979); logistic slopes flatten under unmodelled cluster heterogeneity (Zeger et al., 1988). Current-status data and its nonparametric maximum likelihood are standard (Groeneboom & Wellner, 1992; Jewell et al., 2003), the NPMLE of a mixing distribution is older still (Kiefer & Wolfowitz, 1956; Laird, 1978; Lindsay, 1983), as are the non-identifiability bounds used where budget exhaustion competes with genuine failure (Peterson, 1976; Tsiatis, 1975). What is new is that agentic benchmarks have the design feature that makes the mixture identifiable, the same task run repeatedly against one model, and that nobody had used it.

2.4 Adjacent results

Sinha et al. (2026) shows short-task benchmarks can make progress look slower than it is, and ours is the same category of claim applied to curve shape. Lee (2026a) studies the independent repeated runs our identification depends on, and Iscan (2026) does not confirm an error-content effect in self-repair under a preregistered placebo control — our own test (Appendix F) is a null with stated power, not a falsification, and points the same way. Coston (2026) proposes falsification as an early validity check, the standard our rejected-hypotheses table is written to. Appendix A extends all of this, including the parallel control-task literature.

3 Data and setup

We analyse both public releases of METR's agentic time-horizon evaluation (Kwa et al., 2025), 41,629 runs over 170 tasks in the first release and 24,008 over 228 in the second (METR, 2026), 65,637 in total. Each row is one attempt by one subject at one task, with a binary outcome, the task's rated human completion time, and start and completion timestamps.

Counts differ by filter, so we state the filter at first use. The second release has 21 subjects, 20 AI models and a human control; the first has 34, of which 33 are AI models and 32 clear the estimator's inclusion filter, which keeps cells with at least two runs and requires at least 25 such cells, 8 successes and at least one failure per model; one first-release model fails it. All three task sources are kept and cells enter the likelihood unweighted. Pooled and matched to release dates, 52 models support the trend analysis. Human rows are a control arm and never enter an AI median. **No benchmark model was run for this paper.** Every model name is a column of pre-recorded outcomes in a public log, with no API calls and no inference behind any reported number; the agent session of §6.6 is study method, not data.

Success is current-status data. Write T_j for task j's rated human completion time, n_j for runs and s_j for successes. Treating completion as survival to T_j gives $P(\text{success} \mid T_j) = S(T_j)$, so we

observe *whether* failure occurred before a fixed inspection time and never *when*. That is current-status data, each unit inspected once, and not right-censored survival data; the distinction governs every cross-task estimator that follows. Under a Weibull baseline $S(T) = \exp\{-(T/a)^k\}$, $k < 1$ is a declining hazard and $k > 1$ a rising one, and the published fits report $k < 1$.

Two estimands. *Cross-task hazard* takes T to be a task's rated human completion time, varying across tasks, and measures how success scales with how long a task takes a human. *Within-run hazard* takes T to be elapsed time or tokens generated within a single run, and measures the cause-specific hazard of clean failure among runs still active, with budget exhaustion a competing exit. They are nested, not alternative: within-run, the task-level heterogeneity moves up a level, to a frailty shared by one model's repeated runs of a task, fitted with the same Weibull form (App. B). Within-run outcomes are three-way, success, clean failure and budget exhaustion, and treating budget exits as censoring, a failure time merely unseen, is not innocuous, so we report Peterson bounds on the event probabilities, which hold whatever the dependence between the two exits, and sensitivity refits on the shape (Appendix E.1).

Two further benchmarks, and one validation gate. `monkey_business` (Brown et al., 2024) is a *validation gate* and not a result: its $K = 10^4$ independent samples per problem measure a real heterogeneity distribution where a simulation would assume one. Terminal-Bench (Merrill et al., 2026) trajectories (Lee, 2026b) supply a within-run axis on an independent agent suite (106 agent-model combinations, 51 with token counts logged and so fits on both within-run clocks), and SWE-bench Verified (Jimenez et al., 2024; Chowdhury et al., 2024) trajectories (Bjarnason et al., 2026) supply tool-call count and tokens as primary clocks. Neither receives a cross-task shape estimate: their length covariates are coarse ordinal buckets or unbaselined author estimates.

The estimator is frozen. One implementation, one production grid and one set of tolerances were applied unchanged to every dataset; the grid's history is in Appendix B. The specification, its thresholds and its falsifiers were written down before any extension dataset was touched, in an internal document dated 12 August 2026 and released in the supplementary material. It was not deposited in an external registry and its pre-analysis state is therefore not independently timestamped, so we say pre-specified rather than pre-registered, and deviations carry a date and a reason in the log.

4 Model, identification, and estimation

The model adds one term to the published specification. What this design can identify about that term, and what it cannot, sets the form of every claim in §6.

4.1 Weibull baseline with proportional frailty

Take a Weibull baseline $S(T) = \exp\{-(T/a)^k\}$ and let each task j carry a nonnegative per-task random effect, a *frailty*, $Z_j \sim G$ acting proportionally on the cumulative hazard:

$$S(T \mid Z) \;=\; \exp\!\big\{-Z\,(T/a)^k\big\}\,.$$

Z is the task's unobserved difficulty relative to its rated length and G its distribution across the suite, and a point mass recovers the published estimator. Scale is not separately identified when G is unrestricted, since $Z(T/a)^k = (Za^{-k})T^k$ absorbs any rescaling of a into G; the NPMLE therefore fixes $a = \mathrm{median}(T)$, its only free scalar being k. The naive, gamma and log-normal fits normalise G, to a point mass or to unit mean, and estimate a jointly with k; fixing a in the naive fit, where no G absorbs it, would report 0.32 instead of 0.63.

4.2 Identification comes from repeated runs

What makes G estimable is a design feature agentic benchmarks have and single-run evaluation does not, the same task run more than once against the same model. Conditional on Z_j, the n_j runs of task j are i.i.d. $\mathrm{Bernoulli}(S(T_j \mid Z_j))$, so success counts are overdispersed relative to Binomial for any nondegenerate G; that overdispersion is the information about heterogeneity the marginal curve lacks, and the finite-support mixing distribution is fitted from it. Without repeats nothing separates a hard task from an unlucky one.

Proposition 1 (A fully flexible baseline is not identified)**.** *Under this design — finitely many tasks, each observed at one inspection time T_j with n_j Bernoulli trials — the likelihood depends on the baseline cumulative hazard Λ, here $(T/a)^k$, and on G only through the Laplace transform of G at the points $\{m\,\Lambda(T_j) : m = 1, \ldots, n_j\}$, the chances that m runs all succeed. Nonparametric Λ and G are therefore not jointly identified: any G' whose transform agrees with G's at those points is observationally equivalent. With $n_j = 1$ the degenerate $G = \delta_1$ is always among the equivalent pairs, so single runs cannot detect heterogeneity; with $n_j \geq 2$ it is not, and that overdispersion is what the estimator uses once Λ is restricted to a one-parameter family. Proof in Appendix B.*

On rated length the Weibull restriction is therefore a modelling commitment (App. D.3 relaxes it within-run), and the data corroborate the difficulty: 80.7% and 83.4% of cells in the two releases are degenerate, $0/n$ or n/n, carrying no within-cell information. Two identified relaxations of the form, stratified and piecewise in $\log T$, do not distinguish it from a length-varying elasticity (Appendix B.3). A hazard-ratio argument in Appendix B motivates attenuation but proves no bound on the fitted shape; the exclusion of $k \leq 1$ rests on the calibrated panel analysis of §6.1; the medians reported there summarise per-model fits and are not a replacement value of k.

That exclusion is a statement about the benchmark curve; reading it as a statement about an agent needs one further assumption.

Assumption 1 (Length-invariant heterogeneity)**.** *The distribution G of the task-level frailty Z_j does not depend on the task's rated length T_j.*

Nothing in this design checks it. If residual difficulty scales as a power of length, $Z_j = T_j^{\beta} U_j$, then $Z_j(T_j/a)^k = (U_j a^{\beta})(T_j/a)^{k+\beta}$ identically, so k and β enter only through their sum, and neither replication nor a calibrated placebo separates them (Appendix B.3).

4.3 Estimation, grids and intervals

The four fits differ only in the freedom given to the task-difficulty distribution. The **naive** fit holds G at a point mass and is the published estimator; **gamma** and **log-normal** frailty give G a one-parameter family of unit mean, fitted jointly with k and the scale; the **NPMLE**, the nonparametric maximum likelihood estimate of G, leaves it unrestricted, maximised over mixing distributions on a fixed support grid by EM. That objective is concave in the mixing weights, so EM reaches the global optimum, while the parametric fits are Nelder–Mead with six restarts, are *not* guaranteed global, and were checked against profile likelihood before being called converged.

We use the finite-grid NPMLE because it is frozen, pre-specified and assumes no form for G, and report its recovery, bias and grid sensitivity; the log-normal has lower finite-sample squared error in 9/10 of the simulated conditions we can check and cannot represent the measured heterogeneity exactly (§5).

Grid discipline. Both grids are pre-specified and checked for edge contact on every fit; no fit underlying a headline number contacts an edge.

The support grid is a disclosed sensitivity. Doubling its resolution at fixed span leaves the second-release AI-only median at 1.85 (shift +0.000); widening its span by two decades each side moves 9/20 second-release and 17/32 first-release models by more than one grid step and both medians by less than one, in opposite directions (Appendix B).

Intervals and their calibration. Every quoted interval states its basis. *Profile likelihood* gives within-model uncertainty on k, with simulation coverage of 83–100%; *cluster bootstrap over tasks* gives every interval on a difference between arms and the cross-model reliability of a clock, the share of the cross-model variance in $\hat{k}$ that is not task-sampling noise.

5 What the estimator recovers when the truth is known

Simulation from a set truth shows the sign correction to be a property of the estimator and not of the agents. In simulations of METR's own design — its task set, per-cell run counts and success levels

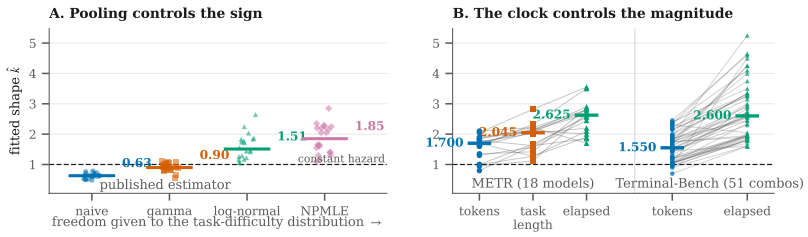

Figure 1: **The two choices. (A)** The same METR data (20 models) under four estimators differing only in the freedom given to the task-difficulty distribution; the published estimator sits below constant hazard, every estimator that models heterogeneity above the published one, and two of the three above constant hazard. **(B)** The 18 models measured on all three clocks, and the same comparison on a benchmark sharing no tasks, agents or harness.

— at fixed true shapes, the naive estimator returns 0.57 at constant hazard under gamma heterogeneity and 0.43 under a heterogeneity distribution measured in an independent dataset and transplanted onto METR's design (below). The NPMLE returns 1.05 under the gamma heterogeneity and 1.150 under the measured one (App. C).

The estimator is biased, and the direction is used. Under each arm's own design at constant true hazard, the NPMLE is biased *upward* on the AI arms — +0.115, positive in 18/18 models — and *downward* on the human arm, returning 0.781 $[0.688, 0.927]$. The upward bias makes every agent point estimate if anything too high, and is the fact behind the realised level of the exclusion test in §6.1; the downward bias means the human cross-task figure is read against 0.781, not against one.

A measured heterogeneity distribution, not an assumed one. Brown et al. (2024) provides 10^4 samples per problem, enough to measure a heterogeneity distribution nonparametrically (127 items, $\mathrm{Var}(Z) = 21.49$), which we transplant onto METR's design before refitting. At $k_{\text{true}} = 1.65$ the estimators return 0.581, 0.911, 1.566 and 1.900. The NPMLE overshoots, and on Wasserstein distance to the measured distribution the log-normal beats it below roughly $K = 100$ samples per item and loses above, winning on finite-sample error at benchmark sample sizes and losing asymptotically because the measured distribution is not log-normal.

A null it must not pass. Shuffling the outcome-to-length assignment within model destroys any real relationship. Across 1,000 such fits the highest shape returned was 1.60, and 20/20 models exceed their own null on the real data. The estimator does not manufacture the effect.

6 Results

6.1 Pooling controls the sign

Letting task difficulty vary reverses the fitted shape's sign (Figure 1A). Across 20 AI models in the second release the median shape is 0.63 under the naive fit, 0.90 under gamma, 1.51 under log-normal and 1.85 under the NPMLE; the first release gives 2.00 across subjects, 2.025 on AI models.

The panel-level exclusion survives calibration. Proposition 1 limits what these data identify without the Weibull restriction, §5 shows the NPMLE is biased upward here, and both cut against per-model certificates (Figure 2). 17 of 20 models have profile intervals strictly above one (counting lower ends at exactly one gives 19), but the realised per-model false-exclusion rate at constant true hazard is 0.111 $[0.087, 0.140]$, not the nominal 0.025. **The per-model exclusions do not carry the certificates they appear to; we do not claim them.**

The count survives even so. All models share one task suite, so their exclusions are dependent, and we simulate the panel at constant true hazard with a shared task draw. Under the most conservative dependence tested, shared frailties with a shared task resample, the count has mean 3.21, a 99th percentile of 13 and a 99.9th percentile of 15; its maximum in 3030 replicates is 16 of 20, and 0/3030 replicates reach the observed 17, which bounds that count's null probability at 1.0×10^{-3}

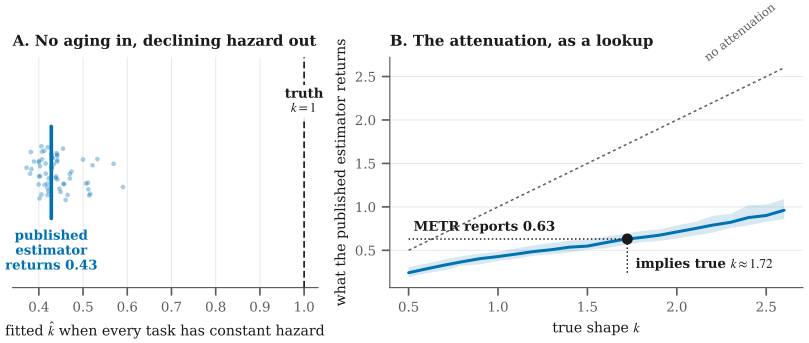

Figure 2: **The attenuation, twice. (A)** METR's real design simulated with every task at constant hazard and the heterogeneity measured from an independent dataset: the published estimator returns 0.43. Nothing in the simulated world is aging, and the estimator reports that agents get safer. **(B)** The same attenuation as a lookup: read at the published 0.63, it implies a true shape of 1.72.

(95% upper bound, rule of three). Dropping the short-task family carrying most of the difficulty pile-up leaves 162 tasks and wider intervals, and moves the median to 1.900, upward, where our bias account predicts a fall; we do not explain it.

The sign survives the aggregation; the magnitude does not. The median above is one aggregation of 20 per-model fits. Pool the same data under one shape and a task difficulty shared across models and the estimate falls to 1.05 $[0.99, 1.12]$, covering constant hazard. The data reject that specification's no-interaction constraint at a likelihood-ratio statistic of 926 (a boundary null, so no p-value is reported). Restoring the model-by-task interaction, the admissible pooled fit returns 1.56 $[1.40, 1.74]$, with constant hazard rejected at a likelihood ratio of 59.6 on one degree of freedom. So across the aggregations that the data do not reject, the corrected shape spans 1.56 to 1.85. The fits quoted are the blind reimplementation's (§6.6); our own fits of both specifications, in Appendix E, agree to within a grid step.

6.2 The clock controls the magnitude

The cross-task axis used above is one of three clocks indexing the same runs, and the choice between them moves the fitted shape by about as much as the pooling correction does (Figure 1B). For the 18 models measured on all three, median $\hat{k}$ is 1.700 on tokens generated (14/18 models above one), 2.045 on task length (18/18; above §6.1's 1.85 because the two models without a tokens clock sit low) and 2.625 on elapsed time (18/18). The ordering holds model by model in 16/18, 16/18 and 17/18 comparisons. The per-model difference elapsed minus tokens averages +0.989 ($p = 0.0003$), and the mean tokens-minus-task-length difference is -0.320 ($p = 0.0024$), which rules out the reading that the split is between within-run and cross-task clocks. Subtracting the medians above does not give these means of per-model differences.

It replicates. On Terminal-Bench, a different agent suite sharing no tasks, agents or harness, 51 of 106 combinations log token counts and so carry fits on both within-run clocks, at 1.550 on tokens and 2.600 on elapsed, mean gap +1.175 (median +1.000), 50/51 positive, $p = 6.5 \times 10^{-10}$. The two benchmarks agree on both medians to within 0.15 and on the mean separation to within 0.19; three harnesses supply 47 of the 51 combinations, so the replication is a result about those harnesses.

Where it does not. SWE-bench Verified does not reproduce the ordering. Its r2e-gym configurations record a run time of their own, and on that clock the median shape, 2.30, sits *below* the tokens value 2.73. Only 3 of 6 configurations run in the METR direction, at a median within-configuration gap of +0.20 against +1.000 on Terminal-Bench — medians of within-unit differences, not differences of the medians just quoted. That axis is a different quantity: the trajectories record model and environment execution per step, and environment execution is a median 0.95% of trajectory time, so what is called elapsed here is model compute, not the wall time between two submission timestamps that METR and Terminal-Bench record. Whether that field carries latency of its own, these logs do

not settle. We count it as a replication in neither direction, and the ordering result rests on METR and Terminal-Bench (App. C).

Three candidate mechanisms, none sufficient. Figure 4 (Appendix D) summarises what each candidate can account for against the observed offset. *Selection under the estimator's own design* predicts the gap with the wrong sign. Simulating each model's own design at its own truth predicts -0.153 against an observed elapsed-minus-task-length offset of +0.669, a residual of +0.821 ($p = 0.00005$), so design accounts for -22.8%. *A power-law reparameterisation* would make the offset deterministic — if tokens accumulate as elapsed$^{\alpha}$ the ratio of fitted shapes should equal α — but α has median 1.111 against an observed ratio of 1.719 (paired $p = 0.0014$), and the two are uncorrelated across models ($r = -0.262$). *Measurement error in the tokens clock* would attenuate its shape, and is the natural candidate given the paper's own mechanism; simulation–extrapolation (SIMEX; Cook & Stefanski, 1994) sizes it under classical, nondifferential error across the 18-model panel. Under linear extrapolation the correction covers a median 19% of each model's elapsed-minus-tokens gap at noise SD $\sigma = 0.6$ on log tokens, large for a directly recorded integer count, leaving a median 0.98 shape units unaccounted; quadratic extrapolation agrees with linear at $\sigma = 0.3$ but is extrapolant- rather than data-determined beyond it (App. D).

The cross-task and elapsed estimates summarise the same runs under different clocks, separated by an offset we cannot account for; they are not independent routes to one conclusion. What corroborates the finding is that the sign survives on clocks whose failure modes differ.

6.3 A HUMAN CONTROL, SAME TASKS AND SAME CLOCK

Measured on the same recorded clock, humans are indistinguishable from constant hazard (Figure 3), at $\hat{k} = 1.00$ $[0.90, 1.30]$ by cluster bootstrap — indistinguishable rather than constant, since the interval admits a mildly rising human hazard. Under its own design at constant true hazard the estimator returns 1.065, and 1.028 on the matched agent designs, so the design asymmetry runs against the contrast below, not for it (Appendix D).

The 41 tasks they skipped have median rated length 300 minutes against 12.62 for the suite, so we refit the agent arm on the 171 tasks the human fit uses, giving 2.225 against 2.625 unrestricted. On those shared tasks, resampling the task set once per replicate and refitting both arms on the same draw gives $\Delta\hat{k} = 1.225$, 95% CI $[0.825,\ 1.725]$, with 0/300 replicates at or below zero.

A least-favourable scenario, not a bound. Grant that the entire elapsed excess is artefact and none of it reaches the human arm. The agent floor is then its tokens value, the clock least exposed to wall-clock time, 1.700, against a human at 1.00, and 14/18 agent intervals still exclude $k \leq 1$ there, as do 38/51 on Terminal-Bench. The agent floor is a full-suite median against a matched-set human figure and humans have no token clock, so this is a scenario, not a bound on differential clock distortion; human elapsed time contains breaks an agent trace does not, and nothing here bounds which arm that favours.

On task length the human arm gives 0.60 $[0.50, 0.70]$. That covariate is built from the successful baseliners' own durations, a construction whose effect on $\hat{k}$ simulation puts within 0.12 in either direction. The larger channel is selection: a task no baseliner completed has no baseline duration, which biases the estimator downward under the arm's design. Inverted through that design's calibration curve (Appendix D) the figure is 0.67 $[0.54, 0.91]$, still below one; the correction pulls the interval's upper end toward one with the point, so it excludes one by less.

6.4 FURTHER RESULTS, IN BRIEF

Per-model rankings depend on the clock: cluster-bootstrap reliability runs 0.63 on rated length to 0.92 on tokens. Terminal-Bench and the budget-exit/Peterson check are in App. C and E; under both extreme codings of budget exits the point estimate stays above one in 17/20 models. Growth over this period moved the curve's location and left its shape flat to the limit the designs resolve: 1.005 of the horizon trend sits in the scale and -0.005 in the shape, the verdict Xing et al. (2026) reach on METR's data under a Rasch model with rising scalar ability.

6.5 What we tested, and what each check established

The clock-offset row lists three explanations of our own; two rows test our own estimator.

hypothesis	what the check established
Rising hazard is self-conditioning	Pre-specified test with controls detects none; 80% power at log-odds 1.52/1.35 (t-based 1.85/1.94)
The within-run rise is the Weibull's doing	Post hoc free-form (Cox) baseline, same per-task effect: rising in 18/18 on elapsed (2.24), 11/18 on tokens (1.48); shape below one without it (0.79, 0.67; App. D.3)
Our estimator is the most accurate	Measured-G transplant: log-normal wins at finite samples and loses asymptotically
Shape covaries with capability	Conditioned on the count of non-degenerate cells the trend's interval covers zero; $\hat{k}$ tracks that count at $r = -0.82$, and simulation excludes design thinness (App. C)
The human figures are unbiased, and the cross-axis gap is a human anomaly	Its own null: 0.781 $[0.688, 0.927]$, below truth at every k; and the gap is panel-wide, the human's, +0.40, below 14 of the 18 agent gaps
The axis offset is design selection, clock reparameterisation or tokens measurement error	Design predicts the wrong sign and accounts for -22.8%; $\alpha = 1.111$ against a required 1.719; SIMEX at generous σ falls short panel-wide (median 19% of the gap)
Pile-up heterogeneity inflates the headline	Dropping the pile-up-heavy family raises the estimate

6.6 A blind reimplementation, and the choice it found

We gave the four specification files and the public METR data to a separate Claude Opus 4.8 session in Claude Code, and nothing else. Appendix G records the one break in the blinding.

The headline reproduces. From the text alone they obtained 0.6310 naive against our 0.63, 1.8456 corrected against our 1.85, and the same exclusion count, 17 of 20; the first release, the human arm and the fixed-scale check agree too (Appendix G). Per-model corrected shapes agree between implementations at Spearman 0.998; naive and corrected shapes correlate at only 0.59 in theirs and 0.57 in ours, so no model's corrected shape can be read off its naive one.

They reproduced the aggregation sensitivity independently. The 1.05 the data reject and the admissible 1.56 that sets the bottom of the range are both fits they ran without access to ours. A support-grid convention the text left open moves their estimate from 1.64 to 1.87, a spread of 0.23, more than any other source of uncertainty they controlled.

7 Discussion

Model the pooling structure. Give the magnitude across every clock the benchmark records. Where only one clock exists, report the shape as a sign with its aggregation range (procedure: Appendix H).

Claims about *how far* agents can go are unaffected, since nearly all growth sits in the location channel; claims about *how reliably* depend on the shape. At fixed task difficulty (Assumption 1), a curve fitted at $k = 0.63$ understates the ratio of the 80%-reliability length to the 50% length by 2.9–3.3$\times$ across the corrected range (§1).

7.1 Limitations

The axis offset has no mechanism. Shape depends on the clock, the dependence replicates, and three candidate explanations fail; reading that as evidence that one clock is correct claims more than we showed. The elapsed clock cannot be cleaned here, two timestamps and no per-step timing, and the discount we apply to Terminal-Bench's disagreeing elapsed-clock trend for sitting on that exposure applies to our own elapsed figures.

The human arm is thin — 58 informative clusters, 66.1% degenerate — and covers the short end, so the contrast is reported on the matched set (§6.3). Human and agent failure are not the same event, and the schema distinguishes neither. Against the worry that failures simply run longer: the human failure-to-success duration ratio, 22–34$\times$, is above the model median in either release, so if that drove rising shape the human arm would show the highest and it shows the lowest. Only 3 first-release models clear the timestamp screen, so the first release cannot test the clock result. Repeated runs of a task are treated as independent given its difficulty; retries or shared scaffold state would inflate the fitted heterogeneity, and nothing here separates the two. The reimplementer and our analysis-code assistance share a model family, so the check tests whether the specification suffices and not whether two systems agree (App. G).

AI USE STATEMENT

Generative AI tools were used to assist with development, debugging, and review of analysis code written to implement author-specified analyses, and for copyediting and language review of author-written manuscript text. The research questions, study design, statistical methods, analytical decisions, interpretation of results, and scientific content were determined by the authors. All code, analyses, results, citations, and final manuscript text were reviewed and verified by the authors.

The blind reimplementation reported in §6.6 was itself performed by a language-model agent. That is a feature of the study design and is described there and in Appendix G.

ETHICS STATEMENT

This work analyses publicly released evaluation logs and involves no human subjects research. The rows labelled human in METR's data are METR's own published baseline timings, collected and released by them; we did not collect timing data, recruit participants, or observe individuals. No personally identifying information appears in any dataset used here, and none is reconstructible from what we release.

Three of the four datasets carry permissive licences: `monkey_business` under MIT, Terminal-Bench trajectories under Apache-2.0, and the SWE-bench artifacts under CC BY 4.0. METR's `eval-analysis-public` repository carried no licence file at the time of writing, so we redistribute none of its data. Our released code fetches it from the original source instead of vendoring a copy, so reproduction stays possible without our making a redistribution claim we are not entitled to make. Our own code and derived results are released under the MIT licence, chosen so the design calculator can be run on other groups' evaluation data without obligation.

The substantive claim of this paper is that a published curve has been read as saying something it does not say. We state it in a form meant to be useful and not accusatory: the analyses we correct were competently executed, the confound was named in public before any of them, and the reason nobody resolved it is that resolving it requires a design feature whose relevance was not obvious. We name the specific analyses because a correction that does not say what it corrects is not checkable.

A reader might take this work as grounds for greater confidence that agents degrade over long tasks, and act on that. The bound on what we show is stated in §7.1: the sign is established across clocks and benchmarks, the magnitude is not, and no claim here supports a specific reliability figure for a specific deployment.

REPRODUCIBILITY STATEMENT

Every estimator, simulation harness and figure in this paper is produced by a script in the released analysis tree, and every number in the text is a macro in a registry that carries its provenance to a stored result or a documented derivation. The four estimators compared in §4 (naive, gamma, log-normal and the NPMLE), the within-run variants used for the clock result, the cluster-bootstrap and joint across-arm bootstrap procedures behind every interval on a difference, the recovery and design-null simulations of §5, the self-conditioning test of Appendix F, and the free-baseline check of Appendix D.3 are each a named script in that tree, together with the design calculator that maps a proposed benchmark design to the shape bias it would induce and a command-line diagnostic (`curveshape_diagnose.py`) that applies the frozen estimator and its grid-edge check to any run-level table. The tree is submitted as supplementary material and is browsable during review at `https://anonymous.4open.science/r/rs-curveshape-D488/`. Consistent with the Ethics Statement, no METR data is redistributed — that repository carried no licence file at the time of writing — so the released code fetches the evaluation logs from their original source and rebuilds every derived table from them.

Where the details live: the model, its identification argument and the limit on it are stated formally in §4 and Appendix B, including Proposition 1 and its proof sketch; the data sources, inclusion filters and the timestamp-missingness screen are in §3 and Appendix D; the grid conventions and their history are in Appendix B; and the calibration basis of every reported interval — profile likelihood or cluster bootstrap — is stated at its first use, per §4.

The freeze is narrow. The multi-dataset extension, the self-conditioning test, the estimator and its grids, and every directional hypothesis in this paper, other than the checks marked as not pre-specified in Appendices C and D.3, were written down, with their thresholds and their falsifiers, in an internal specification document before the corresponding analyses were run. That document was not deposited in an external registry, and its pre-analysis state is therefore not independently timestamped; we describe this work as pre-specified rather than pre-registered for that reason. The specification and its amendment log, the analysis code, the derived results and the reimplementer's own report all ship in the anonymised supplementary material. Deviations from the specification are logged in the document itself with a date and a reason, never applied silently; those that occurred are reported where they arise (Appendices B, C and F).

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Bjarni Haukur Bjarnason, André Silva, and Martin Monperrus. On randomness in agentic evals. arXiv:2602.07150; data at `https://huggingface.co/datasets/ASSERT-KTH/agentic-evals-artifacts`, 2026. CC BY 4.0 license; 10 independent runs per configuration of 500 SWE-bench Verified instances.

Norman Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99, 1974. doi: 10.2307/2529620.

Norman E. Breslow. Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):216–217, 1972.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. monkey_business (Large Language Monkeys: Scaling Inference Compute with Repeated Sampling). `https://huggingface.co/datasets/ScalingIntelligence/monkey_business`, 2024. MIT license; K=10,000 samples per problem; GSM8K, MATH, CodeContests, MiniF2F-MATH.

Neil Chowdhury, James Aung, Chan Jun Shern, Oliver Jaffe, Dane Sherburn, Giulio Starace, Evan Mays, Rachel Dias, Marwan Aljubeh, Mia Glaese, Carlos E. Jimenez, John Yang, Leyton Ho, Tejal Patwardhan, Kevin Liu, and Aleksander Madry. Introducing SWE-bench Verified. `https://openai.com/index/introducing-swe-bench-verified/`, August 2024. OpenAI.

John R. Cook and Leonard A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994. doi: 10.1080/01621459.1994.10476871.

Amanda Lee Coston. Falsifying discriminant validity of predictive algorithms. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26)*, pp. 3105–3128, 2026. doi: 10.1145/3805689.3812310. arXiv:2601.17146.

David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x.

Piet Groeneboom and Jon A. Wellner. *Information Bounds and Nonparametric Maximum Likelihood Estimation*, volume 19 of *DMV Seminar*. Birkhäuser, Basel, 1992. doi: 10.1007/978-3-0348-8621-5.

Piet Groeneboom, Marloes H. Maathuis, and Jon A. Wellner. Current status data with competing risks: Consistency and rates of convergence of the MLE. *Annals of Statistics*, 36(3):1031–1063, 2008. doi: 10.1214/009053607000000974.

John Gurland and Jayaram Sethuraman. How pooling failure data may reverse increasing failure rates. *Journal of the American Statistical Association*, 90(432):1416–1423, 1995. doi: 10.1080/01621459.1995.10476647.

Fergus Hamilton. Peto's paradox and the future of AI agents. https://gushamilton.github.io/lab/2026/01/23/petos-paradox-ai-agents/, January 2026. Updated 3 February 2026.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Bo E. Honoré. Identification results for duration models with multiple spells. *The Review of Economic Studies*, 60(1):241–246, 1993. doi: 10.2307/2297821.

Mehmet Iscan. Form, not content? A preregistered, placebo-controlled evaluation of learned error-conditioned self-repair through prompts and weights in frozen small code models. *arXiv preprint arXiv:2607.12962*, 2026.

Nicholas P. Jewell, Mark J. van der Laan, and Tanya Henneman. Nonparametric estimation from current status data with competing risks. *Biometrika*, 90(1):183–197, 2003. doi: 10.1093/biomet/90.1.183.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world Github issues? In *International Conference on Learning Representations*, 2024.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887–906, 1956. doi: 10.1214/aoms/1177728066.

John P. Klein. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48(3):795–806, 1992. doi: 10.2307/2532345.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3086.

Nan Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978. doi: 10.1080/01621459.1978.10480103.

Wilson Y. Lee. Capable but unreliable: Canonical path deviation as a causal mechanism of agent failure in long-horizon tasks. *arXiv preprint arXiv:2602.19008*, February 2026a.

Yoonho Lee. terminalbench-trajectories: Terminal-Bench 2.0 trajectories. https://huggingface.co/datasets/yoonholee/terminalbench-trajectories, 2026b. Scraped from the tbench.ai leaderboard; Apache-2.0 license; revision 04e8940, accessed August 2026.

Bruce G. Lindsay. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, 11(1):86–94, 1983. doi: 10.1214/aos/1176346059.

Fengyuan Liu, Jay Gala, Nilaksh, Dzmitry Bahdanau, Siva Reddy, and Hugo Larochelle. BRIDGE: Predicting human task completion time from model performance. In *International Conference on Machine Learning*, 2026. arXiv:2602.07267.

Fernando Martínez-Plumed, Ricardo B. C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42, 2019. doi: 10.1016/j.artint.2018.09.004.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces. In *International Conference on Learning Representations*, 2026.

METR. Time horizon 1.1. https://metr.org/blog/2026-1-29-time-horizon-1-1/, January 2026.

Jonas Moss. (Updated) METR's data can't distinguish between trajectories (and 80% horizons are an order of magnitude off). https://www.lesswrong.com/posts/sBEzomgnYJmYHki9T, February 2026. LessWrong, post ID sBEzomgnYJmYHki9T.

Toby Ord. The Lindy Effect. *arXiv preprint arXiv:2308.09045*, 2023.

Toby Ord. Is there a half-life for the success rates of AI agents? *arXiv preprint arXiv:2505.05115*, 2025.

Toby Ord. Hazard rates for AI agents decline as a task goes on. https://www.tobyord.com/writing/hazard-rates-for-ai-agents-decline, February 2026.

Arthur V. Peterson, Jr. Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. *Proceedings of the National Academy of Sciences*, 73(1):11–13, 1976. doi: 10.1073/pnas.73.1.11.

Frank Proschan. Theoretical explanation of observed decreasing failure rate. *Technometrics*, 5(3): 375–383, 1963. doi: 10.1080/00401706.1963.10490105.

Akshit Sinha, Arvindh Arun, Shashwat Goel, Steffen Staab, and Jonas Geiping. The illusion of diminishing returns: Measuring long horizon execution in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.09677.

Anastasios Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22, 1975. doi: 10.1073/pnas.72.1.20.

James W. Vaupel, Kenneth G. Manton, and Eric Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454, 1979. doi: 10.2307/2061224.

Hanwen Xing, Pengyun Wang, BingXu Meng, Kumail Alhamoud, Xiang Li, Jicheng Wang, Xin Yu, Xinyang Han, Xiaomin Li, Philip Torr, and Yuexing Hao. CurveShift: Is agent progress scalar? Separating level from shape. *arXiv preprint arXiv:2608.00355*, 2026.

Zhanfu Yang. Why retrying fails: Context contamination in LLM agent pipelines. *arXiv preprint arXiv:2605.08563*, 2026.

Scott L. Zeger, Kung-Yee Liang, and Paul S. Albert. Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44(4):1049–1060, 1988. doi: 10.2307/2531734.

A Extended related work

A.1 The horizon literature

METR's time-horizon framework (Kwa et al., 2025) summarises a model by the rated human task length at which it succeeds half the time. Most follow-on work studies the *location* of that curve — how fast the half-success length doubles — rather than its *shape*. Moss (2026) estimates a doubling time, finds that the data cannot tell growth trajectories apart, and shows that variation in difficulty among tasks of equal length puts the marginal 80% horizon about an order of magnitude below the typical task's; Xing et al. (2026) separates level from shape, tests ceiling effects and avoids harness confounding on a scaffold-free benchmark, fitting a Rasch model with a difficulty per item and a scalar ability rising across generations; Liu et al. (2026) ports the framework to further benchmarks with a two-parameter item-response model carrying a difficulty and a discrimination per task. Our decomposition in §6 is consistent with all of it: essentially all observed growth over two and a half years is location.

Item-response models of task difficulty. All three papers model task heterogeneity, and none is subject to the pooling artefact this paper describes. The distinction is the scale the difficulty sits on. An item-response model, the standard tool for instance-level difficulty in machine-learning evaluation (Martínez-Plumed et al., 2019), places difficulty and ability on the success-probability scale and estimates where they cross; it yields no hazard, so none reports whether the failure rate rises or falls as a task runs, and none varies the clock. A difficulty shared across models is also the structure a Rasch model imposes, and that structure is one of the aggregations we fit: under the hazard it returns 1.05, the one specification in §6.1 that does not exclude constant hazard, and the data reject its no-interaction constraint at a likelihood ratio of 926. The difficulty that moves the sign is model-by-task and is identified from repeated runs of one task against one model.

Xing et al. (2026) explain the drift of gains toward harder tasks across generations as ceiling compression on the easy ones, a claim about the trend that our decomposition shares. It is not a claim about the sign, and the two are separable here. A ceiling is a relation between one model and the suite and would leave a model far below it with an undistorted curve; task heterogeneity is a property of the suite and spares none. The naive shape sits below one for all 20 second-release models, from 0.48 to 0.78, and simulating each model's own design at constant hazard returns 1.05 with no slope on the count of cells the model neither always passes nor always fails (App. C).

Shape enters through the parametric fits. Hamilton (2026) fits Weibull and log-logistic families and reports declining hazard, and Ord (2026) reports that finding. The fit is to the pooled curve, and neither post reports a within-run axis. The present paper does not dispute their arithmetic; it disputes what a pooled shape parameter measures.

A.2 Why the log-logistic result already pointed here

Both posts report a slight goodness-of-fit edge for the log-logistic over the Weibull, within the noise of a per-model split that is otherwise even. Under the identity of §2 — gamma frailty at unit shape over a Weibull baseline is the log-logistic — that comparison is not a neutral model-selection exercise. A log-logistic fitting better than a Weibull is what a heterogeneous population predicts, and is consistent with one; goodness of fit alone does not establish that heterogeneity generated the data. The information was present and was read as a statement about which curve to draw; a fitted mixture is what turns it into a statement about what the population contains.

A.3 Statistical antecedents

Proschan (1963) traced the decreasing failure rate of air-conditioning systems pooled across aircraft to a mixture of exponential, constant-hazard, distributions, and Gurland & Sethuraman (1995) showed that pooling can reverse increasing hazards too, the mechanism this paper measures. Vaupel et al. (1979) introduced frailty explicitly for heterogeneous populations and gave the pooled hazard under gamma frailty in closed form. Zeger et al. (1988) give the marginal-versus-conditional attenuation of regression coefficients under unmodelled cluster effects, the regression analogue of a pooled shape compressed relative to the individual one.

The current-status observation model and its nonparametric maximum likelihood are standard (Groeneboom & Wellner, 1992; Jewell et al., 2003); the estimator used here is the NPMLE of a mixing distribution, whose consistency, EM computation, and existence and geometry are due to Kiefer & Wolfowitz (1956), Laird (1978) and Lindsay (1983); Groeneboom et al. (2008) give consistency and rates for the maximum likelihood estimator under current status with competing risks. Where budget exhaustion competes with genuine failure, we use the non-identifiability bounds of Peterson (1976) and the non-identifiability result of Tsiatis (1975) rather than assuming independent censoring.

A.4 Adjacent claims about evaluation artefacts

Sinha et al. (2026) shows that a benchmark's choice of task length can make progress look like it is slowing: marginal gains in single-step accuracy compound into exponential gains in the length of task a model completes, so a short-task suite understates the trend. That is the same category of claim as ours, applied mainly to the level of the curve. The same paper reports self-conditioning, per-step accuracy falling once a model's own errors are in its context, the mechanism Appendix F

tests. It is the closest methodological precedent for the reporting rule we propose. Coston (2026) proposes falsification as an early validity check, the standard the rejected-hypotheses table in §6 is written to. Hewitt & Liang (2019) and Adebayo et al. (2018) established, in probing and in saliency respectively, that a method's findings must be checked against a control in which the signal has been destroyed: a probe should score far lower on random control tasks, and a saliency map should change when the model's weights or the labels are randomised. The permutation null in §5 is the analogue here.

Lee (2026a) studies independent repeated runs of long-horizon tool-use tasks, the design feature our identification depends on, and its increasing prevalence in agentic evaluation makes the estimator applicable beyond this one dataset. Yang (2026) studies retries that keep the failed attempt in context, which break that independence, and attributes the gap seen under independent retries to correlated task difficulty. Iscan (2026) tests self-repair with a placebo-controlled preregistered design; our own falsification test of a related mechanism (Appendix F) is a null with stated power, not a falsification, and points the same way.

B Estimator details

B.1 Likelihood and EM

With task j contributing s_j successes in n_j runs at inspection time T_j, and frailty $Z_j \sim G$, the marginal likelihood is

$$L(k, G) \;=\; \prod_j \int \binom{n_j}{s_j} \left[e^{-z\mu_j(k)}\right]^{s_j} \left[1 - e^{-z\mu_j(k)}\right]^{n_j - s_j} \, dG(z), \qquad \mu_j(k) = (T_j/a)^k.$$

For fixed k this is a mixture likelihood, linear in G, so maximisation over G on a fixed support grid is a concave problem solved by EM to the global optimum. We profile over k on the pre-specified grid and take the maximiser, with the 1.92 log-likelihood drop giving the profile interval.

The EM update runs in probability space with row-maximum scaling: each row of the log-kernel is shifted by its own maximum before exponentiation, the offset accumulated separately, and the weights updated multiplicatively. The mathematically identical logsumexp implementation is roughly forty times slower and changes no estimate to machine precision; the choice is computational, not numerical.

For the within-run axis the cluster $c = (\text{model}, \text{task})$ contributes D_c clean-failure events and per-run clocks t_i, and the z-integral has the same structure with kernel $z^{D_c} \exp\{-z M_c(k)\}$, $M_c(k) = \sum_i (t_i / T_{\text{med}})^k$. The baseline-hazard product depends on k, not on G, and is added to the profile, not dropped — omitting it pins the profile at the grid floor, a silent failure mode.

B.2 Proof sketch of Proposition 1

Write $L_G(u) = \int e^{-uz}\, dG(z)$ and $\Lambda_j = \Lambda(T_j)$. Expanding $[1 - e^{-z\Lambda_j}]^{n_j - s_j}$ binomially, task j's likelihood is a finite linear combination of $L_G(m\Lambda_j)$, $m = s_j, \ldots, n_j$, so the data constrain (Λ, G) only through $N = \sum_j n_j$ values of one function. (i) Rescaling to $(c\Lambda,\ \text{law of } Z/c)$ leaves every $L_G(m\Lambda_j)$ unchanged. (ii) Holding Λ fixed, the vector $(L_G(m\Lambda_j))_{j,m}$ is a G-average of the bounded continuous curve $z \mapsto (e^{-m\Lambda_j z})_{j,m}$ on $[0, \infty)$, so it lies in that curve's closed convex hull and a discrete G' with finitely many atoms reproduces it exactly (Carathéodory), differing from G unless G is already that law. (iii) The homogeneous alternative. For $n_j = 1$ the pairs $G = \delta_1$, $\Lambda_1 = -\log \pi$ and $G = \text{Exp}(1)$, $\Lambda_2 = 1/\pi - 1$ match the only observable, π_j, the second by the exponential's Laplace transform $\int_0^\infty e^{-z\lambda} e^{-z}\, dz = 1/(1+\lambda)$, so heterogeneity is undetectable. For $n_j \geq 2$ they do not: at $\pi_j = 1/2$ they give $P(s_j = 2) = 1/4$ and $1/3$, and in general $E[e^{-2\Lambda_j Z}] > (E[e^{-\Lambda_j Z}])^2$ for any nondegenerate Z, which is the overdispersion of §4. Repeated runs therefore exclude a homogeneous suite; they do not identify Λ. Restricting Λ to the Weibull family with scale fixed by the normalisation in §4 leaves one free scalar, k, whose recovery §5 checks by simulation.

The standard hazard-ratio argument motivates, but does not prove, a downward relationship: the marginal hazard factors as $h_{\text{marg}}(t) = h_{\text{indiv}}(t) \cdot E[Z \mid T > t]$, and $E[Z \mid T > t]$ is nonincreasing in t for any G with finite mean, so the marginal log-hazard's local slope in $\log t$ sits everywhere

weakly below the individual curve's. That is a statement about a local slope, not about the fitted Weibull shape under the binomial current-status loss used here — a projection onto finitely many inspection times T_j, not a derivative — so we do not claim the naive fit's $\hat{k}$ is a proven lower bound on the individual shape whatever G is. The exclusion of $k \leq 1$ therefore follows from replication, not from this argument; the two must not be merged.

B.3 Why a length gradient in the frailty is untestable

Assumption 1 cannot be checked from data of this shape. The obstruction is an exact equivalence, not a shortage of power. Let residual difficulty scale as a power of rated length, $Z_j = T_j^{\beta} U_j$ with $U_j \sim G_U$ free. Then for every task,

$$Z_j \left(\frac{T_j}{a}\right)^k \;=\; U_j\, T_j^{\beta}\, a^{-k} T_j^{k} \;=\; \left(U_j a^{\beta}\right) \left(\frac{T_j}{a}\right)^{k+\beta},$$

so the pair (k, β, G_U) and the pair $(k+\beta,\ 0,$ law of $U_j a^{\beta})$ imply identical success probabilities in every cell — not approximately, but as the same model written two ways. This is Proposition 1's non-identification one level up: with G left free, β is absorbed into the shape at zero likelihood cost.

Two remedies suggest themselves and neither works. The first leans on the repeated runs that identify G. Replication within a task pins that task's success probability and hence the *spread* of G across tasks; a length gradient moves G's *location*, and under the equivalence above the location at the fitted shape is constant in T for every β. A band-wise comparison of the fitted G's location across task-length strata therefore targets a quantity that is identically zero under both the null and every alternative. Run on this data, it returns a value that tracks where the support grid ends, flipping sign when those endpoints move with the grid's resolution held fixed.

The second is a placebo null, and it fails more completely. Calibrating one by simulating from each model's own fitted pair and comparing the real statistic against it is circular: when the alternative (k, β, G_U) holds, the fitted pair *is* $(k+\beta, 0,$ law of $U_j a^{\beta})$, so the simulated "null" data are pointwise identical in distribution to data generated under the alternative. Power equals size exactly, at every β.

What survives is the reading, not the estimate. The fitted shape is the benchmark curve's fall-off against rated length, and the identity above says a length-scaling difficulty gradient belongs to that fall-off; it does not contaminate it. Only the agent-level gloss — that a given agent's hazard rises within a task of fixed difficulty — requires Assumption 1. A gradient that is *not* a power of length is identified, as lack of fit against a baseline left piecewise in $\log T$. Two identified relaxations of the frozen form, a baseline stratified across 5 length bands and a two-segment piecewise baseline with G free, both return model-resample intervals on the local-elasticity contrast that cover zero, $[-0.17, 1.43]$ and $[-0.50, 0.63]$, and the two constructions agree with each other at chance (sign agreement 12/20). That is not evidence the Weibull is true; the data do not distinguish a length-varying elasticity from the frozen form.

B.4 Grids, and how they were widened

The k grid runs from 0.30 to 4.00 in steps of 0.05; the frailty support grid is logarithmic with 170 points running from 10^{-5} to 10^{5}. Both were widened after an earlier stage of this work reported estimates from a k grid capped at 2.51, at which a quarter of fits sat on the boundary. Every fit reported here is checked for edge contact and none underlying a headline number touches one.

Three support-grid conventions have existed across scripts in this project: the production grid above, an undocumented one-decade-wider drift to 10^{-6} to 10^{6} that several later robustness scripts inherited, and a deliberately wide 10^{-9} to 10^{9} used only for a boundary-mass audit. That audit refits every subject inside the competing-risks code with a warm-started, shorter EM, and its production-grid fits sit one grid step from the headline fits for 8 of 21 second-release subjects; within the audit's own fits the AI-only median moves $1.875 \to 1.850$ on the second release and $2.000 \to 2.050$ on the first, inside one grid step either way. On the production fit path, doubling the grid's resolution at fixed span moves the second-release median by +0.000, and the reimplementation of Appendix G, on a support two decades wider each side, moves it by +0.015. All headline numbers use the production convention.

Separately, the blind reimplementation of Appendix G, working from a decade count without endpoints, ran its own narrow, production and wide sensitivity check over 10^{-4} to 10^{4}, 10^{-5} to 10^{5}, and 10^{-6} to 10^{6}, each holding the point count fixed, finding the estimate runs 1.64 to 1.87 across them, a spread of 0.23 against the sub-grid-step figure below. The endpoints above let a reader reproduce either check without guessing at the range; the sensitivity of the median to the alternatives, on our own convention, is reported in §4 and is smaller than one grid step on the second release.

B.5 Interval calibration

Profile-likelihood intervals attain 83–100% coverage across the simulated truths in §5. Cluster bootstrap resamples tasks with replacement, recomputes the scale normalisation within each replicate, and drops replicates with fewer than four events — a guard against non-convergence that also truncates the distribution it is estimating, so we report that it did not fire on the contrast: the 300 replicates behind the interval below are 300 draws retained, all 18 models fitted in every one. For any difference between two arms, one task draw is shared by both arms within a replicate. That is what makes an interval on a difference honest; it is not what makes it narrow. Over 300 replicates the two arms correlate at $r = -0.05$, and the realised standard deviation of the difference, 0.250, sits just above the 0.246 that independent arms would give — so the shared draw buys correctness here, not precision.

C Validation and multi-dataset detail

C.1 Recovery on the real design

We simulate METR's own design — its task set, per-cell run counts and success levels — at fixed true shapes and refit. The naive estimator returns 0.57 at $k_{\text{true}} = 1$ under gamma heterogeneity, 0.69 at 1.5, and 0.43 at $k_{\text{true}} = 1$ under the heterogeneity measured in an independent dataset and transplanted onto this design (App. C.2; interquartile range 0.407–0.458 over 60 replicates on 223 cells). Under gamma heterogeneity at the same three true shapes 1, 1.5 and 0.63, the NPMLE returns 1.05, 1.60 and 0.65; under the transplanted heterogeneity at $k_{\text{true}} = 1$ it returns 1.150. A model's *informative cells* are the task–model cells it neither always passes nor always fails, $0 < s_j < n_j$; the rest carry no within-cell information. The slope of recovered shape against that count is indistinguishable from zero at every tested truth level, so the estimator's behaviour does not depend on how thin a particular model's design is — which matters because the count correlates with the estimate itself at $r = -0.82$ in the real data (Spearman -0.92). Simulation excludes design thinness as the cause of that correlation. We do not offer a positive account of it.

Under difficulty pile-up — a mass of near-impossible items, which is what the measured G contains — recovery at $k_{\text{true}} = 1$ is 1.050 and 1.150 at two pile-up levels, both a level shift above truth rather than a slope (paired t: $p = 0.0032$; Wilcoxon $p = 0.0051$). This level shift is why the exclusion of $k \leq 1$ has realised level 0.111 rather than 0.025.

Figure 2 in the main text shows this attenuation twice: METR's real design at constant hazard, and the same relationship as a lookup.

C.2 The measured-G transplant

Brown et al. (2024) provides 10^4 samples per problem, enough to estimate a problem-difficulty distribution nonparametrically rather than assume one. We measure G there (127 items, $\mathrm{Var}(Z) = 21.49$) and transplant it onto METR's design.

At $k_{\text{true}} = 1.65$ the four estimators return 0.581, 0.911, 1.566 and 1.900 respectively. The NPMLE overshoots. Wasserstein error against the measured G is 0.049 at low true shape and 0.007 at high, against 0.014 and 0.110 for the log-normal, with the crossover between $K = 50$ and $K = 100$ samples per item. A beta-binomial reference plateaus at 0.015–0.022.

The log-normal wins on finite-sample error at the sample sizes benchmarks have and loses asymptotically, because the measured G is not log-normal. In 9/10 of the simulated (sample size, true shape) conditions the log-normal has lower error, of which 4 have margins below one grid step and should not be counted as wins.

C.3 TERMINAL-BENCH

Terminal-Bench supplies 106 agent-model combinations. On elapsed time the median shape is 2.70 with 106/106 above one; on tokens 1.55 with 45/51 above one among the 51 combinations with fits on both clocks. The filter is not an edge rule in practice: no Terminal-Bench fit on either clock touches a grid edge, and every excluded combination is one whose harness logged no token counts, so no excluded combination has a tokens fit that could have reversed the ordering. The included set is concentrated, three harnesses supplying 47 of the 51, and skews toward combinations with more recorded failures; the ordering result on this benchmark is about those harnesses. Cluster-bootstrap reliability is 0.88 and 0.91. Cross-axis correlation within combination is 0.658 (Spearman 0.723), falling to 0.55 when independent units are enforced — markedly higher than METR's 0.10–0.24.

Terminal-Bench also shows a within-run shape trend against release date that excludes zero. We report it and decline to read it as capability covariation, for three reasons: it does not replicate on the same dataset's tokens axis, its information-content control removes far less of the raw association than METR's does, and the calendar span is under one year, so the per-year rate extrapolates well beyond the data.

C.4 SWE-BENCH VERIFIED

SWE-bench Verified puts the shape above one on every clock and replicates the clock ordering in neither direction. On the nano-agent trajectories, tool calls give 1.92 and tokens 2.15. On the r2e-gym trajectories — six configurations, three models each at two decoding temperatures, ten runs per item — the recorded run time gives 2.30, tokens 2.73 and steps 3.90, with 6/6 configurations above one on each. Extending from two to ten runs per item moves the estimate by a median -0.85, down in all 14/14 configuration–clock cells. Two runs per cluster resolve the frailty distribution poorly and the NPMLE sits high there, the direction of the upward bias §5 measures on METR's own design; the low-replicate estimates overstate the magnitude, and the sign, 6/6 configurations above one at ten runs, does not move.

The recorded run time sits *below* tokens here, 2.30 against 2.73, and 3 of 6 configurations show the elapsed-minus-tokens difference the other two panels do, at a median within-configuration gap of +0.20 against a median +1.000 on Terminal-Bench. Those are medians of within-unit differences, not differences of the medians quoted above. The two configurations furthest from the METR direction are the devstral pair, and those are also the only two whose failures are *shorter* than their successes on that clock; 4/4 of the remaining configurations run in the ordinary failures-longer direction. Failures concentrated early in the run flatten the hazard on that clock, which is what a low shape there is, so the low shape and the duration anomaly are one fact.

The clock itself is why this dataset does not test the ordering. These trajectories record model execution and environment execution per step, and the two sum to the recorded total in 720 of 720 instances checked, drawn from the first run of each configuration. Environment execution is a median 0.95% of trajectory time and at most 4.2% by the same per-configuration median, with harness test evaluation accounted separately and comparably small, so the recorded run time is model compute. We did not refit on a compute-only clock: a per-run rescaling of a few percent cannot move a shape by the gaps at issue, and the decomposition bounds the effect more tightly than a refit would. METR and Terminal-Bench record wall time between two submission timestamps, which carries queueing, retries and idle, the spread §6.2 measures as a median 19.7$\times$ interquartile ratio of tokens per elapsed minute within a model. The two clocks are different quantities. On a clock with no queueing in it, the reading under test — that wall-clock exposure is what separates elapsed from tokens — predicts no excess, and none appears. That removes the dataset as evidence against the reading and supplies no mechanism for the METR offset, which stays open. The decomposition was run after the reversal had been seen, to test whether it could be a contaminated-clock artefact, and is not part of the pre-specification.

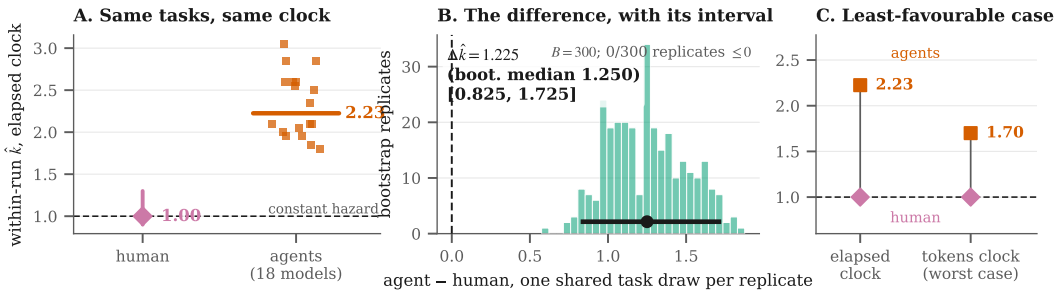

Figure 3: **The human control. (A)** Both arms on the elapsed clock, restricted to the 171 tasks both arms have fitted data on, every agent drawn rather than only the median. **(B)** The difference, from a bootstrap resampling the shared task set once per replicate and refitting both arms on that draw; the panel marks the bootstrap median, 1.250, and the text reports the point contrast 1.225, with the same interval. **(C)** The sensitivity check of §6.3: a least-favourable assumption, not a bound.

D THE WITHIN-RUN AXIS, AND THE CLOCK RESULT

D.1 CONSTRUCTION AND SCREENING

Elapsed time is `completed_at` minus `started_at`, both epoch milliseconds. Tokens is `tokens_count`, null for human rows, so the human control exists on the elapsed clock alone — a structural asymmetry that constrains what the clock result can say about the contrast, and the reason for the least-favourable scenario in §6.3.

Timestamps are not missing at random. In the first release, 57% of successes lack them against 5.6% of failures; a naive complete-case analysis there would be conditioning on the outcome. We screen models to below 5% per-model missingness before fitting and report the differential. Only 3 first-release models survive that screen, against 18 in the second, which is why the clock result is a second-release and Terminal-Bench finding and why the first release cannot test it. That is a coverage limitation of the release, not a failed replication.

D.2 THREE CLOCKS

Median shape over the 18 second-release models measured on all three clocks: 1.700 on tokens, 2.045 on task length, 2.625 on elapsed. Per-model, tokens is below task length in 16/18 comparisons, task length below elapsed in 16/18, tokens below elapsed in 17/18. Elapsed minus tokens averages +0.989 across models ($p = 0.0003$); tokens minus task length averages -0.320 ($p = 0.0024$). Both are means of within-model differences; the three figures above are medians across models.

On Terminal-Bench, 51 of 106 combinations log token counts and so carry fits on both within-run clocks: tokens 1.550, elapsed 2.600, gap +1.175 (median +1.000), 50/51 positive, $p = 6.5 \times 10^{-10}$. Intervals excluding $k \leq 1$: 38/51 on tokens, 51/51 on elapsed.

D.3 THE SIGN WITHOUT THE WEIBULL

This check was not pre-specified. It asks whether the Weibull form carries the sign, and it can run only on the two within-run clocks, where each run's failure or censoring time is observed. On rated task length a run contributes one pass or fail at a single length, and Proposition 1 shows that a free baseline and a free G are then not jointly identified; Appendix B fits the relaxations that are. On the within-run clocks, repeated runs of each task identify both, up to scale (Honoré, 1993). We refit the 18 agent models; the human arm and Terminal-Bench keep the Weibull.

We keep the per-task effect, with G unrestricted, and replace the Weibull baseline with the Cox model's (Cox, 1972), estimated as in Breslow (1972; 1974): a cumulative hazard that jumps at each observed failure time and commits to no shape, as an empirical CDF replaces a fitted distribution. Two summaries put the free fit on the paper's scale. k_{free} is the slope of log cumulative hazard against log time over the middle 80% of failure times, since the extremes rest on few failures; for a Weibull

this slope is exactly k. R splits that window at the median failure time and divides the mean failure rate in the later part by the rate in the earlier part, so $R > 1$ means a rising rate. Its size grows with the span of the window, so R is read against one and not compared across clocks. R decides the sign and k_{free} gives the magnitude. G sits on 677 log-spaced support points over the main estimator's span: a free baseline flattens the likelihood, and the main estimator's coarser grid biases k_{free} low by up to 0.08. Intervals come from a task bootstrap ($B = 250$) with one task draw per replicate, shared across models and clocks. A model is called rising when its interval on R lies above one, falling when it lies below, and undetermined otherwise. These are nominal 95% intervals, not corrected across the 18 models; in simulation with a flat true hazard the call said rising in 3 of 50 datasets.

Rise-then-fall hazards matter here, because four models below show one on the token clock. In simulations on this design, with Weibull, Gompertz and rise-then-fall baselines under gamma or log-normal G, the Weibull fit called a rise-then-fall hazard rising in 28% of replicates under gamma G. The free fit's R fell on the correct side of one in at least 97% of replicates in every cell where the true hazard was not flat, and the mean error of k_{free} stayed below 0.03 in every cell. A second implementation with a different optimiser agrees with ours to within 0.006 in k_{free}.

clock	Weibull k	k_{free}	R	calls	no per-task effect
elapsed	2.625	2.24 [2.09, 2.56]	46.1 [20.0, 89.8]	18 / 0 / 0	0.79 [0.70, 0.93]
tokens	1.700	1.48 [1.36, 1.63]	2.32 [1.64, 3.16]	11 / 4 / 3	0.67 [0.59, 0.74]

Table 1: Medians over the 18 models with a free-form (Cox) baseline and the same unrestricted per-task effect, with 95% task-bootstrap intervals. Calls count models rising / falling / undetermined from each model's interval on R. The last column is k_{free} from the same free-form baseline with the per-task effect removed.

On elapsed time the Weibull's 2.625 becomes a k_{free} of 2.24 [2.09, 2.56], and all 18 models are called rising (Table 1). With a gamma G in place of the unrestricted one (Klein, 1992) it is 1.97 [1.81, 2.16]. Three specifications on one clock span 1.97 to 2.625, so the baseline and the mixing distribution both move the magnitude. Without the per-task effect the same free-form fit has k_{free} below one on both clocks, 0.79 on elapsed time and 0.67 on tokens: the published statistic, from a fit with no Weibull in it. Its R is below one on tokens, 0.41 [0.28, 0.51], and undetermined on elapsed time, 1.37 [0.87, 2.31]. Per-model k_{free} on elapsed time exceeds its tokens value by a mean of 0.86 [0.68, 1.27], in 16/18 models (Weibull: +0.989, 17/18). Widening the window to the middle 90% of failure times, or narrowing it to the middle half, moves the elapsed-time median to 2.39 or 2.02 and the tokens median to 1.53 or 1.43; every elapsed-time fit keeps R above one, and 12 or 13 token fits do.

On tokens the free baseline keeps the median sign: k_{free} is 1.48 [1.36, 1.63] and R is 2.32 [1.64, 3.16]. Model by model the token clock was never uniform: 14/18 Weibull fits sit above one (§6.2). Under the free baseline 11 models rise, 3 are undetermined and 4 fall: GPT-4 (November 2023), GPT-4o, Claude 3 Opus and Claude 3.5 Sonnet (June 2024), all released before July 2024. Their token hazards are humped, low over the first few thousand tokens, highest between about 3k and 130k tokens, and lower after that. The fall begins far below these runs' token budgets of about 8M, and recoding budget exits as failures leaves R below one for all four. A single Weibull exponent averages across the hump and lands near one, between 0.80 and 1.30. All four rise on elapsed time. We do not pursue the cause of the hump here.

D.4 Three candidate mechanisms and why each fails

Selection under the design. Simulating each model's own task-length design under its own inclusion rule at three truths — constant hazard, its own task-length estimate, its own elapsed estimate — gives estimator bias of +0.115 at $k_{\text{true}} = 1$, positive in 18/18 models. The gap this predicts is -0.153 against an observed +0.669, leaving a residual of +0.821 ($p = 0.00005$). Design explains -22.8% of the offset: it points the other way.

The same simulation run on the human arm gives 0.781 $[0.688, 0.927]$ at $k_{\text{true}} = 1$ — biased *downward*, because a task no baseliner completed has no baseline duration and is coded from an estimate, a selection on at least one success; without that selection the same design returns 1.13. The human

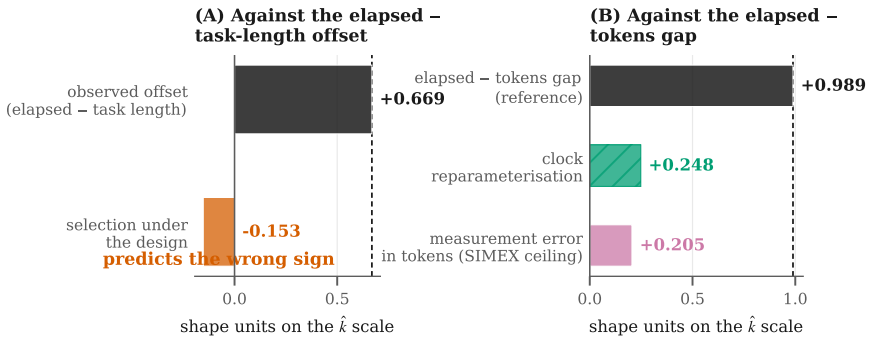

Figure 4: **What each mechanism can account for**, on the same additive scale as the observed offset. Selection under the estimator's own design predicts the offset with the wrong sign. Neither of the other two candidates reaches half of it. **(A)** The observed offset and the design-selection bar, read against the elapsed-minus-task-length offset, +0.669. **(B)** The reparameterisation and SIMEX bars, read against the elapsed-minus-tokens gap, +0.989; each panel carries its own reference line. Bars are panel means of the absolute correction each candidate delivers (SIMEX: linear extrapolation at $\sigma = 0.6$); the 19% and quarter-of-the-offset figures in the text are medians of per-model ratios.

cross-task figure of §6.3 is read against this null by inverting the calibration curve $E[\hat{k} \mid k_{\text{true}}]$ that the same simulation traces over k_{true} from 0.47 to 1.25. The two arms' designs bias in opposite directions, which is why the human's cross-axis gap, +0.40, looked anomalous before each subject was compared to its own null and unremarkable after: it is below 14 of the 18 agent gaps.

Clock reparameterisation. If tokens accumulate as elapsed$^{\alpha}$ then a Weibull in one clock is a Weibull in the other with shape scaled by $1/\alpha$, and the ratio of fitted shapes should equal α. Regressing log tokens on log elapsed within model gives α with median 1.111 and per-model standard errors under 0.08, against an observed shape ratio of 1.719: paired $p = 0.0014$, and the two are uncorrelated across models ($r = -0.262$). The elasticity accounts for about a quarter of the offset and none of its cross-model structure.

Measurement error. Attenuation of the tokens clock would be the most parsimonious account, since attenuation under covariate error is the mechanism the whole paper rests on. SIMEX adds noise of known SD σ to log tokens, refits at each noise level and extrapolates the shape back to zero noise. Across the panel, 17 of 18 models have a positive elapsed-minus-tokens gap (one is excluded, its gap negative), and the panel's gaps span -0.15 to $+1.90$ with median 1.10. We judge each model's correction against its own gap rather than a panel-wide correction, since the two are not comparable in absolute terms. Under linear extrapolation, coverage has median 19% (IQR 10%–34%) at $\sigma = 0.6$ — large for a directly recorded integer count — leaving a median 0.98 shape units of the typical model's gap unaccounted, essentially unchanged across $\sigma \in [0.3, 1.5]$; only 2 of 17 models are fully covered even at $\sigma = 1.5$. Quadratic extrapolation agrees with linear at $\sigma = 0.3$ (both $\approx 11\%$) but diverges 7-fold by $\sigma = 1.5$; beyond $\sigma \approx 0.6$ the quadratic answer is extrapolant- rather than data-determined. A handful of models with especially small gaps make the fraction numerically unstable there, so we report medians, never means, and do not treat the quadratic figures beyond $\sigma = 0.3$ as informative. This bound assumes classical, nondifferential measurement error on tokens; it does not hold, in either direction, if the error is differential or Berkson-type. For the model examined before the panel existed, whose elapsed estimate exceeds its tokens estimate by +1.15, the correction at $\sigma = 0.6$ is +0.138 linear and +0.447 quadratic, close to the panel's own median. The implementation reproduces the frozen estimator exactly at zero added noise, and the attenuation is monotone in the added noise for every model checked, so the machinery behaves; the magnitude is not there for most of the panel.

What is not in question. Within a single model, the interquartile ratio of tokens per elapsed minute has median 19.7×, ranging 7.4–108.5× across models. Wall-clock time on this benchmark does not measure work done. The released schema has two timestamps and no per-step timing, so the two cannot be separated within it. Erratic throughput also tracks the elapsed excess across

models ($r = 0.60$, $p = 0.009$); we report that as consistent-with rather than evidence-for, since both sides of the correlation contain elapsed time.

D.5 Human arm, in full

The human arm has 769 runs over 187 tasks with 215 clean failures and zero administrative censoring; after the pre-specified two-run minimum, 753 runs over 171 tasks, 66.1% degenerate, 58 informative clusters. Shape 1.00, cluster bootstrap $[0.90, 1.30]$.

Humans baselined 187 of 228 tasks. The 41 they skipped have median rated length 300 minutes against 12.62 for the suite. Refitting the agent arm on the 171 tasks the human fit uses gives 2.225 against 2.625 unrestricted; the joint bootstrap on the shared task set gives $\Delta\hat{k} = 1.225$ $[0.825, 1.725]$ over 300 replicates, with 0/300 at or below zero.

The elapsed-clock design null. The cross-task calibration above has an elapsed-clock counterpart. Simulating the human arm's own design at $k_{\text{true}} = 1$, with $\hat{G}$ fitted at that truth, each run's latent clean-failure time drawn from the cause-specific hazard and its observed end time kept as the censoring time, gives 1.065 (Monte Carlo SE 0.014, 5th–95th percentiles 0.85–1.35, 100 replicates); drawing censoring times from the arm's censored runs instead gives 1.068. The same simulation on each of the 18 matched agent designs returns a median 1.028 (range 1.002–1.053, positive in 18/18, 30 replicates each). Both arms are biased upward on this clock and the human arm more so, so the contrast the two designs alone produce at a common truth is -0.04; subtracting each arm's bias moves the human figure to 0.94 $[0.84, 1.24]$ and the point contrast from 1.225 to 1.26. Two limits: keeping the observed end times as censoring times can only shorten the simulated design, which carries a median 0.70 of the real event count on the agent arms and 0.82 on the human arm; and the agent correction applies a bias measured at $k_{\text{true}} = 1$ to estimates near 2.2, so the calibrated contrast is a first-order reading, not a refit.

Human failure durations exceed human success durations by 22–34×, a ratio of medians, above the model medians of 4.5× and 9.6× in the two releases, though individual cells span 2.3–159× across 52/52 cells. If that asymmetry drove rising shape the human arm would show the highest, and it shows the lowest.

E Robustness

E.1 Budget exhaustion as a competing risk

A run that stops because it exhausted a token or time budget has not failed in the sense the model assumes, and treating those exits as ordinary censoring assumes independence that nothing here establishes. Budget exits are 9.2% of runs and 24.4% of failures in the first release, 12.4% of failures in the second.

We therefore report Peterson bounds, which assume nothing about the dependence between the two causes. They bound the event probabilities; the refits under the two extreme codings are sensitivity analyses on the fitted shape, not derived bounds on it. Under the most adverse coding the AI-only median falls to 1.23 in the first release and 1.40 in the second, and 21/32 and 17/20 models respectively keep a point estimate above one under both extreme codings (11/32 and 11/20 keep the lower profile bound above one under both). In 15/20 second-release models the bracket is narrow enough that the assignment of budget exits does not change the reading at all.

Bracket width correlates with budget share ($r = 0.93$ in the second release, 0.47 in the first), as it should: the bound is wide exactly where the ambiguity is. Budget exits also ramp with elapsed time in 28/31 first-release and 4/8 second-release models, counted over the models that record budget exits at all. That ramp makes independent censoring implausible and the bounds necessary.

Remark 1 (Direction of the budget-exit shift)**.** *Recoding a budget exit from failure to censoring removes an event from the numerator at a late clock value while leaving the risk set intact, which lowers the estimated hazard on that late interval. That is a local statement at fixed G; it does not by itself fix the direction of the profiled shape once G is re-maximised on the recoded data, and we do not prove one. A three-state fit — success, clean failure and budget exit as competing exits with*

cause-specific Weibull hazards over one shared task frailty — estimates the clean-failure shape k_1 directly, a different estimand from the headline. That fit is consistent with the local direction. Of the models for which the three-state fit converges, 12 in the first release and 7 in the second, the shape falls in 17/19, with median shifts of -0.35 and -0.60 in the two releases, and 5/7 and 8/12 of those models still exclude one afterwards.

E.2 Era-varying caps

Time and token caps changed over the evaluation period, so a failure near the cap means something different early and late. Recoding failures within a tolerance of the operative cap as censored, era by era, and refitting the trend analysis changes no sign and no stated conclusion.

E.3 Sensitivity of the within-run estimate to budget handling

Recoding budget-exit runs as clean-failure events, rather than censoring them, moves the pooled second-release within-run estimate from 2.600 to 2.600 — unchanged — and the first release from 1.675 to 1.750; 6/19 second-release models shift by more than 0.10. The pooled figures here are not the AI-only headline and should not be compared to it directly.

E.4 Measurement error in the task-length covariate

The cross-task covariate is a rated human completion time, itself estimated. Calibrating the error from METR's own repeated human baselines gives a median log-scale standard error of 0.446 in the second release and 0.387 in the first. SIMEX correction moves the naive estimate by +0.031, uniformly upward across 10/10 cells — the expected direction, since covariate error attenuates. The NPMLE moves by -0.009 under linear and -0.030 under quadratic extrapolation, with 3/6 cells moving up, 2/6 moving the wrong way and 1/6 flat. The correction is small relative to the effect and has no consistent direction under the corrected estimator.

On the human arm, SIMEX gives 0.62–0.68 against its own naive fit of 0.38, leaving the human cross-task figure below one under every correction we can apply.

E.5 Task-level frailty as an alternative specification

Fitting frailty at the task level rather than the model level gives 1.02 $[0.95, 1.12]$ on the second release across all subjects and 1.03 on AI models, and 0.96 $[0.91, 0.98]$ on the first. These sit below the per-model estimates and near one.

specification forces a single shared task difficulty across all models, and the log-scale intensity still correlates with each model's own shape at -0.54 and -0.33 in the two releases. Relaxing the pooled shape alone, while still sharing one task frailty across models, closes only 4–9% of the gap to the per-model headline; the tension is instead a model-by-task interaction, which the data prefer to the shared-frailty fit at a likelihood-ratio statistic of 926 against a boundary null. Our own profile of the pooled shape over that interaction's variance, on the 20-model specification, gives 1.55 $[1.40, 1.74]$, a likelihood ratio of 59.6 against constant hazard, matching the reimplementation's 1.56 $[1.40, 1.74]$; the other two specifications were not profiled. Per-model shape spans a factor of 2.6–2.9, so the constraint is a real oversimplification. The reading is that under a specification forcing one shared task difficulty the pooled estimate sits near constant hazard, and that this is a property of the constraint.

E.6 Permutation null

Shuffling the outcome-to-length assignment within model and refitting, across 1,000 fits at 50 permutations per model (on the drifted support grid of Appendix B), the null sits at the k-grid floor: per-model median 0.30, 95th percentiles up to 1.40, the highest single draw 1.60. 20/20 models exceed every draw of their own null on the real data.

E.7 Growth over the record: scale, not shape

Across 52 models spanning two and a half years, 1.005 of the observed horizon trend sits in the scale and -0.005 in the shape. Conditioning on that decomposition, the shape trend is +0.056/yr $[-0.006, 0.174]$, against paired within-model noise worth 9.7 years of it — flat to the limit this design resolves, which is why nothing in the record forced anyone to look at what sets the shape.

F The self-conditioning falsification test

F.1 The hypothesis and the design

If within-run hazard rises because an agent that has erred becomes more likely to err again — recency-structured self-conditioning, which Sinha et al. (2026) report in long-horizon execution — then conditioning on a recent error should predict the next error above and beyond the run's general difficulty. We pre-specified a test of that mechanism before running it, with its decision rule, its negative controls, and a mandatory positive control, on nano-agent SWE-bench trajectories and on Terminal-Bench.

The design is case-crossover in structure: each run serves as its own control, so anything constant within a run — task difficulty, model identity, scaffold — cannot confound the comparison. It has two arms. Arm A compares the exposure history at a run's terminal, failure-ending step against earlier eligible steps within that same run. Arm B instead conditions on survival to a fixed landmark step and asks whether exposure observed up to that landmark predicts the run going on to terminate in failure. The exposures are tool-level failures the agent can see in its own transcript: on nano-agent a shell command exiting non-zero (E1) and a failed `apply_patch` (E2); on Terminal-Bench an error marker in the observation (E1) and a command repeated identically (E2). Negative-control exposures are events with no plausible mechanism linking them to subsequent failure — on Terminal-Bench, using `grep`, reading a file, an observation longer than the median, and a listing or navigation command. The positive control is an exposure that must show an effect if the design has any power at all.

We pre-specified a null expectation. The decision rule required calibration of each observed statistic against its own negative-control distribution rather than against a nominal threshold, which is what makes the resulting minimum detectable effect interpretable.

F.2 Result

The verdict is a null with stated power, not a falsification.

Of the cells, 3 of 4 returned the pre-specified null pattern and 0/4 were positive. Calibrated z-scores were -0.52 and -3.90 on nano-agent and -1.39 and -1.64 on Terminal-Bench — all at or below zero, none in the direction the hypothesis predicts. The positive control fired at +1.952, so the design has power. Sign agreement across arms was 6/6.

The landmark analysis asks whether hazard jumps at the specific points where self-conditioning would act or rises smoothly through them; it gives -0.12 — no landmark structure.

Cap-driven exits, which would manufacture an apparent effect, occur in 0/6 nano-agent and 0/70 Terminal-Bench cells.

F.3 What the minimum detectable effect is, and what it is not

Effects are differences in log-odds between the high- and low-recency exposure bands, so the positive control's +1.952 is a factor of about 7 on the odds. The minimum detectable effects are 1.52 on nano-agent and 1.35 on Terminal-Bench: at two-sided $\alpha = 0.05$ and 80% power, calibrated against each dataset's own negative-control null, this is the smallest true Δ the design would catch four times in five, not an upper confidence limit on the true effect. Both nulls are built from few negative controls, so the thresholds are themselves imprecise: a χ^2 95% interval gives plausible ranges $[0.93, 3.63]$ and $[0.76, 5.02]$, and a t-based calculation at these few degrees of freedom gives 1.85 and 1.94. Effects below roughly 1.3 were outside this design's reach on either dataset.

The design had 80% power to detect effects of 1.52 on nano-agent and 1.35 on Terminal-Bench, and detected none. We do not claim the mechanism is absent: it could be real at a magnitude smaller than these thresholds, which this design cannot see.

F.4 Deviations from the specification

Terminal-Bench's Arm A bootstrap was run first at 200 replicates, under compute pressure, then rerun at the specified 500; every interval and z-score was essentially unchanged between the two runs, a precision confirmation rather than a revision. Four negative-control exposures were used on Terminal-Bench rather than the specified six, because two of the specified controls turned out to be unavailable in the released trajectories. Terminal-Bench's landmark arm (Arm B) was run at landmark $L = 10$ only and without the per-run configuration covariates nano-agent's own Arm B adjusts for. A timed test showed that the full configuration-level dummy encoding needed by this dataset's primary-rung population — the task-length band its primary cells are drawn from — did not converge cleanly and would have made the bootstrap intractable at the specified replicate count. Terminal-Bench's Arm B bootstrap was reduced under compute pressure to 100 replicates and then, after a memory-leak incident during the first attempt, to a final 30; point estimates do not depend on the replicate count, only interval precision does. All are logged with their dates and reasons in the specification document.

F.5 Relation to Iscan (2026)

That paper tests a related self-repair mechanism with a placebo-controlled preregistered design and does not confirm a content-attributable effect; its author reads those results as a failure to confirm, not a refutation. Our null points the same way on different data, a different mechanism and a different exposure definition. Neither result rules its mechanism out.

G A blind reimplementation from the specification

§6.6 summarises what an implementer working only from §3, §4, Appendix B and Appendix E obtained on the public METR releases. This appendix gives the comparison in full, and the places where our text underdetermined the implementation.

G.1 Blinding

The implementer was Claude Opus 4.8 (`claude-opus-4-8`) running in Claude Code, in a separate session with no memory of this project and no access to our code, outputs, figures or manuscript, given the four specification files and the public data. It worked in a two-core sandbox with network access used only to clone that data and with no web search. It shares a model family with the assistance used on our own analysis code, so the check tests whether the specification suffices and not whether two independent systems agree. An agent blinds more completely than a human replication usually does, because it cannot recall what it has not been shown, and less completely in one respect: reading a specification, it settles an ambiguity by picking a plausible reading instead of asking. That is what makes its ambiguity log worth reading.

The blinding was imperfect in one way. A comment header in one of the four files carried an editorial note naming a direction for the support-grid sensitivity. The implementer flagged it and reports having fixed its own grid convention before running that check. Its report ships unedited in the supplement, including that disclosure.

G.2 What matched

quantity	ours	reimplementation
naive shape, second release, AI arm	0.63	0.6310
corrected shape, second release, AI arm	1.85	1.8456
models excluding $k \leq 1$	17 of 20	17 of 20
naive shape, first release	—	0.7007
corrected shape, first release, AI arm	2.025	2.0191
human control, naive	0.37	0.3759
human control, corrected	0.60	0.597
naive shape with the scale held fixed	0.32	0.3184

Per-model corrected shapes agree between the two implementations at Spearman 0.998 on the 20 second-release AI models, with a largest per-model difference of 0.038. Within the reimplementation, naive and corrected per-model shapes correlate at only 0.59 (0.57 in ours), so the correction is not a rescaling of the naive fit. The agreement is on the estimates, not on the intervals: the reimplementation reports profile intervals only and did not attempt the cluster bootstrap, so every interval we quote on a difference between arms is unreplicated.

G.3 What did not, and why that is the result

The reimplementation fitted one specification we describe but do not headline, in which a single shape is pooled across models over a task difficulty shared between them. Without the model-by-task interaction, that fit returns 1.05 $[0.99, 1.12]$, which does not exclude constant hazard. The constraint that produces it is rejected by the data at a likelihood-ratio statistic of 926, and restoring the interaction returns 1.56 $[1.40, 1.74]$ with constant hazard rejected at 59.6. Two readings of that are available and we hold both. The sign is robust to how the panel is aggregated, since the only aggregation that reaches constant hazard is one the data reject. The magnitude is not, since it runs from 1.56 to 1.85 across aggregations we would each have defended.

G.4 Where the specification was underdetermined

The implementer logged twenty-three points at which our text did not fix the implementation. Four of them move a number.

The task filter. §3 fixes it; the specification did not, and the implementer resolved it by working backwards from our stated span. Keeping all three task sources reproduces the length span we report; dropping the short-task family would cut it to roughly three decades and remove most of the leverage on the shape. The implementer identified this as the single most consequential open choice.

The support-grid convention. Appendix B states the endpoints. A decade count and a point count without them leave three conventions open, and under the implementer's reading the estimate runs 1.64 to 1.87 across those, a spread of 0.23, exceeding every other source of uncertainty they could control. Our own sensitivity moves in the opposite direction over what we called the same range, which is itself evidence that a decade count does not fix a grid.

The inclusion filter. §3 states the rule where it gives the count. Given the count alone, the implementer could not recover a filter that removes any model.

Task weighting. The released data ship equal-task and inverse-root-task weights. Our likelihood is an unweighted product of binomials, which §3 says in words and the implementer inferred correctly from Appendix B.

G.5 Two failure modes

Both are silent, both would corrupt a corrected estimate without crashing, and neither is specific to this dataset.

The mixture EM has zero as an absorbing state. Warm-starting from a neighbouring grid point, together with a simplex projection that clips an extrapolated weight to exactly zero, permanently kills support points; the profile then acquires discontinuities at the points where a cold restart happens to run, and per-model maximisers cluster at artificial values. The fix is to cold-start every grid point and never clip a weight to zero.

The alternating sum in the Gamma Laplace transform loses digits with cell size. For the human control arm's largest cells the greatest value the kernel can take falls below the rounding error of the sum, so the whole row is uninformative and the profile acquires holes. The fix is to route such cells to a log-space quadrature.

H THE REPORTING RULE, AS A PROCEDURE

Applied to a benchmark that records repeated runs per task and at least one length covariate, the rule of §7 is five steps.

1. Fit the naive homogeneous Weibull and the same Weibull with a task-level mixing distribution G estimated from the repeated runs (§4). Check every fit for grid-edge contact before reading it.
2. Take the sign from the corrected per-model shapes against a joint null simulated on the benchmark's own design with a shared task draw (§6.1). Report the count of models excluding $k \leq 1$ and what the null produced; do not report per-model certificates.
3. Refit under every aggregation the data do not reject and record the shapes as a range. Three are the minimum: per-model fits summarised by the median; a pooled fit with a model-by-task interaction; and a pooled fit with a single task difficulty shared across models, the structure an item-response model imposes. Test the third against the second by likelihood ratio and drop it from the range only when the data reject it, as they do here (§6.1).
4. Refit on every clock the benchmark logs (tokens, rated length and elapsed time where present), and record the span.
5. Print the sign as a finding and the magnitude as the range across aggregations and clocks. Where the benchmark logs one clock, print the sign with the aggregation range and say the clock is single.

On METR the rule prints: sign, $k > 1$, with 17 of 20 models excluding $k \leq 1$ against a joint null whose maximum in 3030 replicates was 16; magnitude, 1.56 to 1.85 across aggregations and 1.700 (tokens) to 2.625 (elapsed) across clocks, with rated length at 2.045; and, under Assumption 1, an 80%-to-50% reliability-length ratio of 0.484 to 0.542 against 0.165 at the published shape. The command-line diagnostic in the released tree (`curveshape_diagnose.py`) performs the two fits and the edge check on any run-level table with a task-level length covariate; the within-run clocks need the within-run likelihood of Appendix B and are not part of it.