

# Three Hazards Under One Horizon: Constant-Hazard, Degrading-Step and Log-Logistic Accounts of Agent Failure Describe Different Levels of the Same Data, the 50 Percent Trend Barely Depends on Which Is Right, and Every High-Reliability Forecast Does

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

**Abstract.** Forecasts of autonomous AI capability increasingly rest on one number: the length of task, measured in human working time, that an agent completes half the time. That number has doubled roughly every seven months, and three accounts of why agents fail on longer tasks now sit behind its extrapolation. One fits a logistic curve in the logarithm of task length; one proposes a constant failure rate per minute of human task time, giving each agent a half-life; one reports that per-step accuracy falls as a run proceeds, because models condition on their own earlier errors. They are usually read as rival models of one process. This paper argues that they describe three different levels of the same data - the hazard inside a run, the hazard per unit of a task's length, and the success curve across a heterogeneous task suite - and that the aggregation linking those levels is the classic problem of unobserved heterogeneity, under which a falling pooled hazard is compatible with constant or rising hazards inside every task. Published evidence at each level points in different directions: rising, flat and cliff-shaped within-run failure all appear, depending on model, harness and task. Two consequences follow, one reassuring and one not. The 50 percent trend is nearly indifferent to which account is right, since every account is calibrated to it. Every forecast at higher reliability is not. Holding the 50 percent horizon fixed, curves that all match the published 80 percent ratios place the 99 percent horizon anywhere from about one hundredth to about one eight-hundredth of it, and the within-task hazard a deployment would face, which those ratios do not identify, could put it as high as one eighth. A reading procedure for horizon claims and six measurements that would settle the open part are offered. The analysis is built from published measurements, each credited to the source that reported it.

**Keywords:** AI agents; time horizon; capability forecasting; survival analysis; hazard rate; unobserved heterogeneity; long-horizon execution; evaluation methodology

SONYTECH ORCID: 0009-0003-7224-029X

## 1. Introduction

The most quoted measurement in recent AI forecasting is a duration. Kwa et al. (2025) timed skilled humans on a suite of software, machine-learning and cybersecurity tasks, ran AI agents on the same tasks, and for each agent estimated the human task length at which the agent's success rate falls to one half. They called this the 50 percent time horizon, reported that it had doubled approximately every seven months since 2019, and noted that extrapolating the trend would put agents able to complete month-long software tasks within about five years (Kwa et al., 2025). The metric was quickly adopted beyond its authors. It has been reproduced from model performance data alone, with a doubling time of about six months (Liu et al., 2026); folded into a model of compute and algorithmic progress that projects how slowdowns in compute would delay it (Whitfill et al., 2025); audited as one of the few longitudinal series in frontier-AI forecasting (Costa, 2026); and decomposed into a level and a shape over task difficulty (Xing et al., 2026). A survey of 1,547 papers on long-horizon agents ends by naming, as an open problem, whether long-

horizon reliability admits a general predictive theory at all (Chen et al., 2026).

The metric came with a curve, and the curve came with an implicit theory of failure. Kwa et al. fit each agent's success rate as a logistic function of the logarithm of human task time, with the 50 percent horizon as its midpoint and an agent-specific slope. Seven weeks after the first version appeared, Ord (2025) proposed a simpler reading of the same data: a constant rate of failure for each minute a human would take, so that success decays exponentially with task length and each agent can be characterised by a half-life. On that reading, the 80 percent horizon should be a fixed fraction,  $\ln(0.8)/\ln(0.5)$ , or about 0.32, of the 50 percent horizon, and each further "nine" of reliability should divide the horizon by roughly ten (Ord, 2025). Months later, Sinha et al. (2025) reported a controlled result that the constant-rate reading cannot accommodate: when models execute a long sequence of simple steps with the plan and knowledge supplied, their per-step accuracy degrades as the sequence lengthens, partly because a model whose context contains its own earlier errors becomes more likely to err again. They called this self-conditioning. A failure process that worsens with its own

history has no constant hazard to extrapolate.

The natural reading of these three papers is as rival models of one process, of which at most one can be right. Framed that way, the field holds three mutually inconsistent models of how agents fail over steps, and the self-conditioning result breaks any extrapolation built on a constant or independent-step assumption. This paper argues that the natural reading is mistaken in a specific way, and that correcting it changes what can be said about extrapolation.

The argument has four parts. First, the three accounts are not about the same quantity. The logistic fit is a measurement model across tasks, in which human time plays the role of item difficulty; the constant-hazard account is a claim about failure per unit of a single task's length; the self-conditioning result is a claim about failure per step inside a single run. Section 3 states each exactly and Section 4 shows how they connect. Second, the connection between the levels is the problem survival analysts call unobserved heterogeneity, or frailty: when units differ in their failure rates, the hazard of the pooled population falls with time even if every unit's hazard is constant or rising (Proschan, 1963; Vaupel et al., 1979). Ord (2025) himself noted this possibility for METR's task suite. It means that the shape of the cross-task curve cannot, by itself, reveal the shape of the within-task process. Third, the evidence at the within-run level is itself split. Rising hazards are reported for non-thinking models on synthetic execution tasks (Sinha et al., 2025; Mittal, 2026), flat per-round failure for a multi-agent questioning game (Potash, 2026), stable accuracy irrespective of injected errors for thinking models (Sinha et al., 2025), and cliffs for some models on some tasks (Mittal, 2026; Shojaee et al., 2025). The hazard looks like a property of a model, a harness and a task together, which harnesses can change (Meyerson et al., 2025; Pai and Xian, 2026).

Fourth, and most consequential for forecasting: the disagreement matters very little for the 50 percent trend and a great deal for everything else. Every account is calibrated to pass through one half at the measured horizon, so the doubling time of that horizon does not depend on which is right. What does depend on it is any statement about reliability above one half, and any statement about tasks much longer than the horizon. Sections 3.4 and 7 show, by derivation from the functional forms, that holding the 50 percent horizon fixed, curves consistent with the published 80 percent ratios disagree about the 99 percent horizon by a factor of about eight, that two curves agreeing exactly at 50 and 80 percent can still disagree by a factor of nearly two at 99 percent, and that the within-task hazards reported in the literature widen the range further. Kwa et al. state plainly that their data cannot measure horizons at very high success rates (Kwa et al., 2025). The reliability horizons that deployment decisions need are therefore not measured but assumed, and the

assumption is the one in dispute.

Two objections should be met at the start. The first is that this is a statistical quibble about a metric whose authors already hedge it. They do hedge it, about external validity, task distribution and future changes in the trend (Kwa et al., 2025). The hedge this paper adds is different in kind: it concerns which forecasts the data can support even on the measured task distribution, if the trend continues. The second objection is that the self-conditioning result plainly breaks extrapolation, so the right conclusion is that horizon forecasting is unfounded. The evidence does not support so strong a conclusion. The same paper that reports self-conditioning reports that thinking models do not self-condition (Sinha et al., 2025), the trend in the 50 percent horizon survives a change in the failure model, and the qualitative transcripts in Kwa et al. show the failure mode most like self-conditioning becoming rarer as models improved. What breaks is narrower and more specific: the inference from the measured 50 percent horizon to horizons at reliabilities the measurement never reached.

What follows is the author's analysis of published work: it consolidates the reported measurements, derives what each failure model implies, and credits every reported measurement to the source that reported it.

## 2. Scope, exclusions and how the evidence was read

### 2.1 What the paper covers

The subject is the relation between models of agent failure and extrapolations of task-length horizons. The paper covers the METR time-horizon programme and its direct descendants (Kwa et al., 2025; Rein et al., 2025; Wijk et al., 2024; Liu et al., 2026; Whitfill et al., 2025), the accounts of failure that have been offered for its data (Ord, 2025; Sinha et al., 2025), and the 2025-2026 evidence on how agent success varies with the length of a task, the number of steps in a run, and the history inside a run.

### 2.2 What it excludes, and why

Three neighbouring questions are excluded or kept short. First, the general question of whether step errors in multi-step reasoning are independent, whether failure is absorbing, and whether every step is load-bearing was treated in an earlier record in this journal (Mahendrakar, 2026), which used Kwa et al., Ord and Sinha et al. only for what they say about dependence between steps and explicitly left horizon extrapolation "to separate treatment". This paper is that treatment. Where it needs the earlier paper's conclusions, chiefly that step-error independence has no support and that recovery is regime-bound, it states them briefly and refers back rather than re-arguing them. Second, the external validity of METR's task suite - whether software tasks scored automatically stand for real work - is discussed by

its authors and by others (Kwa et al., 2025; Becker et al., 2025; Kapoor et al., 2026) and is used here only where it interacts with the failure model. Third, the econometrics of the calendar trend - what drives the doubling, and whether compute growth will continue - is excluded except where it bears on what the failure models imply (Whitfill et al., 2025).

### 2.3 How the evidence was read

The load-bearing sources were read in full text, listed in the disclosure. Kwa et al. (2025) exists in four arXiv versions, and some of its numbers changed between the first version, which Ord (2025) analysed, and the current one. Where a number differs between versions, the version is named. Several of the most relevant 2026 sources are single-author preprints that have not been peer reviewed (Mittal, 2026; Potash, 2026; Costa, 2026). They are used because they are the only published measurements of the quantities at issue, and their status is stated where they carry weight.

## 3. Three accounts, stated exactly

Much of the apparent conflict between the three accounts dissolves, and the part that remains sharpens, once each is written down in its own terms. This section does that. Throughout,  $t$  is a task's length in human working time,  $h_{50}$  is an agent's 50 percent horizon, and  $S(t)$  is the probability that the agent succeeds on a task of length  $t$ .

### 3.1 The logistic fit is a measurement model with time as difficulty

Kwa et al. (2025) convert every agent run to a binary outcome, rate every task by the geometric mean time of successful human baselines, and fit, separately for each agent, the regression

```
p(success) = sigma( beta_agent * (log h_agent - log t_task) )
```

where  $\sigma$  is the logistic function and  $h_{\text{agent}}$ , the 50 percent horizon, and  $\beta_{\text{agent}}$  are learned parameters. They describe the method as a relative of item response theory: like IRT it finds the difficulty at which an agent has an even chance of success, but unlike IRT it takes difficulty directly from human time rather than learning it from agent performance (Kwa et al., 2025). Read as a survival function in  $t$ , this is the log-logistic distribution (Bennett, 1983), a point Ord (2025) makes when he notes that the curve only looks logistic on a logarithmic axis.

Two features of the fit matter here. First, it is a model of how success varies across different tasks, each attempted by the agent from scratch. It says nothing about when inside a run the agent fails. Human time enters as a difficulty score, and the fit would be unchanged if the tasks' times were the same but every failure happened in the first minute. Second, the slope  $\beta_{\text{agent}}$  fixes how

sharply success falls around the horizon, and therefore fixes the ratio of the 80 percent horizon to the 50 percent horizon: setting the expression equal to 0.8 gives  $h_{80} = h_{50} * 4^{(-1/\beta)}$  when the logarithm is natural. A slope of one gives a ratio of one quarter.

The 50 percent horizon has a property worth stating once, because Section 7 relies on it. Its doubling time is estimated by regressing the logarithm of each agent's horizon on release date. In the current version this gives 207 days, with a 95 percent bootstrap interval of 166 to 240 days; the first version gave 212 days, with 171 to 249 (Kwa et al., 2025). The authors report that the uncertainty in one model's horizon moves together with that of the others, since which tasks of a given human length happen to be sampled shifts every model's estimate in the same direction, so they are "more confident in the slope of the time horizon trend than in the time horizon of any particular model" (Kwa et al., 2025). The trend is a statement about the midpoints of many curves, not about their shapes.

Two later analyses make the IRT reading explicit. Liu et al. (2026) fit a two-parameter logistic IRT model to agent outcomes across many benchmarks, show that the learned item difficulty is approximately linear in the logarithm of human completion time, and use that relation to infer human times for unannotated benchmarks, reproducing a doubling of roughly six months. Xing et al. (2026) show that, on METR's data, a single Rasch model with steadily rising ability reproduces the apparent shift of gains towards harder tasks, which they attribute largely to ceiling effects. In both, human time is a coordinate on a difficulty scale. Neither paper treats it as exposure to a risk.

### 3.2 The constant-hazard account treats time as exposure

Ord (2025) reads the same curves as survival curves. If an agent faces a constant probability of failing in each minute a human would take, then  $S(t) = \exp(-\lambda t) = 2^{-(t/h_{50})}$ , and  $h_{50}$  is the agent's half-life. The account has three attractions that Ord states. It has one parameter instead of two. It explains why the doubling times of the 50 and 80 percent horizons were nearly equal in the first version of Kwa et al. (212 and 213 days), since a halving of the hazard doubles every horizon at once. And it supplies a mechanism: tasks "involve increasingly large sets of subtasks where failing any one fails the task" (Ord, 2025).

Its sharpest prediction is the reliability ladder. Under a constant hazard the  $p$ -horizon is  $\ln(p)/\ln(0.5)$  times the 50 percent horizon, so the 80 percent horizon is 0.322 of it, the 90 percent horizon about one seventh, the 99 percent horizon about one seventieth, and each further nine divides by about ten (Ord, 2025). Ord compared the first prediction with the one published ratio then available: Claude 3.7 Sonnet's 80 percent horizon of about 15 minutes against its 50 percent horizon of 59

minutes, a ratio of 0.25, which he judged "close to this theoretical estimate and within the margin of error" (Ord, 2025; Kwa et al., 2025, v1). He also overlaid fitted exponentials on Kwa et al.'s per-model curves and reported that they fit some models better and others worse, and the data "roughly as well overall".

Ord was explicit about the limits of the claim. He did not assert a precisely constant rate; he proposed it as a baseline from which deviations, a rising or falling hazard, would be informative. He called the result "merely suggestive", asked for a formal comparison of the exponential against the log-logistic fit, and asked whether it applied beyond the task suite (Ord, 2025). He also raised the caveat that becomes central in Section 4: if some tasks are inherently easier per unit time than others, the suite-level curve averages exponentials with different rates, which "produces an aggregate decay curve with thicker tails than an exponential", so that a constant hazard could hold for every task without being visible in the aggregate (Ord, 2025). An earlier paper by the same author develops the mathematics of such mixtures in another setting (Ord, 2023).

### **3.3 The degrading-step account counts steps, and its own formula assumes what its data refute**

Sinha et al. (2025) work in steps rather than minutes. Their opening proposition is the step-level version of Ord's model: if a model succeeds at each step independently with constant probability  $p$  and cannot self-correct, the number of steps at which it reaches success rate  $s$  is  $H_s = \ln(s)/\ln(p)$ , so that the horizon grows hyperbolically as per-step accuracy approaches one, and a small gain in per-step accuracy can buy a large gain in task length. The authors present the proposition as illustrative, made under two assumptions their experiments then test. The arithmetic itself is old: sequential prediction, in which later inputs depend on earlier outputs, was shown to break the independence assumptions of statistical learning in imitation learning (Ross et al., 2011); errors in sequence models were described as accumulating along the generated sequence (Bengio et al., 2015); and multi-step compositional tasks were argued to decay rapidly with complexity in autoregressive generation (Dziri et al., 2023). What Sinha et al. add is a measurement of how the per-step rate itself behaves.

The experiments refute the first assumption. With the plan and the required knowledge supplied in the prompt, and a task requiring only retrieval and addition at each turn, models with near-perfect first-step accuracy still lose task accuracy quickly; the best of the Qwen3 and Gemma3 models in that experiment, Qwen3-32B, fell below 50 percent within 15 turns. Per-turn accuracy itself fell as the turn count rose (Sinha et al., 2025). To separate two explanations, they replaced the model's chat history with synthetic histories carrying a chosen error rate. With an error-free history, accuracy at turn

100 was still below its starting level, which they attribute to long-context degradation; as the injected error rate rose, accuracy at turn 100 fell further. That second effect is self-conditioning. Scaling to frontier non-thinking models of over 200 billion parameters largely removed the long-context component over 100 turns but not the self-conditioning component; the paper's figure caption states that scaling increased it (Sinha et al., 2025).

Two further results from the same paper bear directly on extrapolation. Qwen3 models with thinking enabled did not self-condition: their turn-100 accuracy stayed stable regardless of the error rate in their history, and their thinking traces did not refer back to earlier answers. And removing prior turns from a non-thinking model's context, a form of context engineering, mitigated the effect (Sinha et al., 2025). The degrading-step account is therefore not one claim but two: that per-step accuracy can fall with history, and that whether it does depends on the model's training and on what its harness keeps in context.

### **3.4 What each account predicts, side by side**

The three accounts can be placed in one family for comparison. The Weibull survival function  $S(t) = \exp(-\ln 2 * (t/h50)^k)$  passes through one half at  $h50$  for every shape  $k$  (Weibull, 1951). Its hazard is constant when  $k$  equals one, rising when  $k$  exceeds one and falling when  $k$  is below one. The log-logistic form of Section 3.1 has a hazard that falls monotonically when its slope is at most one. Table 2 sets out what each shape implies for the horizons at higher reliability.

Three things follow from Table 2 by arithmetic alone. First, the 80 percent ratio is a diagnostic of shape. A rising hazard pushes it above 0.32, a falling hazard below. Second, the published 80 percent ratios fall below 0.32. The first version of Kwa et al. gave 0.25 for one model; the current version describes models' 80 percent horizons as "4-6x shorter" than their 50 percent horizons, a ratio of one sixth to one quarter (Kwa et al., 2025). Read as survival curves, these imply a falling hazard: a Weibull shape between about 0.63 and 0.82, or a log-logistic slope between about 0.77 and 1. Because the ratios are read off fitted log-logistic curves, they are partly a property of that family: it is the fitted slope that fixes them, and a log-logistic curve with a slope of about 1.22 would have matched Ord's ratio of 0.322 with a hazard that first rises and then falls. Third, two forms calibrated to the same 50 and 80 percent horizons - the log-logistic with slope one and the Weibull with shape 0.82 - still disagree at 99 percent by a factor of about 1.8, and across the four falling-hazard rows, all inside the published band, the 99 percent horizon runs from one ninety-ninth to one eight-hundred-and-sixth of the 50 percent horizon. The 80 percent ratio constrains the shape near the middle of the curve. It does not constrain the tail. The rising-hazard row is excluded at the level of

**Table 2.** Reliability horizons implied by six failure shapes, as fractions of the 50 percent horizon. Each entry is derived from the stated functional form. The last column is the number of doublings of the 50 percent horizon needed before the 99 percent horizon reaches the same length, assuming the ratio stays fixed.

| Shape                    | Hazard   | h80/h50 | h90/h50 | h99/h50 | Doublings to close 99% gap |
|--------------------------|----------|---------|---------|---------|----------------------------|
| Weibull $k = 2$          | rising   | 0.567   | 0.390   | 1/8.3   | 3.1                        |
| Exponential (Ord)        | constant | 0.322   | 0.152   | 1/69    | 6.1                        |
| Log-logistic, slope 1    | falling  | 0.250   | 0.111   | 1/99    | 6.6                        |
| Weibull $k = 0.82$       | falling  | 0.250   | 0.100   | 1/177   | 7.5                        |
| Log-logistic, slope 0.77 | falling  | 0.167   | 0.058   | 1/380   | 8.6                        |
| Weibull $k = 0.63$       | falling  | 0.167   | 0.051   | 1/806   | 9.7                        |

the suite by the published ratios; Section 4 explains why it is not excluded inside a task.

Figure 1 draws the same comparison. Panel (a) shows four survival curves, each passing through one half at the horizon; they differ most far from the horizon, which is where forecasts about very long tasks and very high reliability live. Panel (b) shows the p-horizon ratio for each, with the published 80 percent values marked.

The first lesson of this section, then, is that the published 80 percent ratios do not sit where Ord's constant hazard puts them; in the current version of the source they sit clearly below. The second, which the next section develops, is that a ratio below the constant-hazard value is exactly what a constant hazard inside each task would produce in the aggregate, if tasks differ in their rates.

## 4. Three levels, not three rivals

### 4.1 The levels

The three accounts become compatible, and their remaining disagreement becomes precise, once three levels are kept apart.

**Level A: the run.** Inside one attempt at one task, an agent takes steps. At each step there is some probability that it makes an error from which the run does not recover. The within-run hazard is that probability as a function of how far the run has gone and what has happened in it. Sinha et al. (2025) and Mittal (2026) measure at this level, on synthetic tasks in which every step can be checked.

**Level B: the task family.** Hold the kind of task fixed and vary its length - more rows to process, more files to edit, more turns of the same operation. The within-family curve is success probability as a function of length. Ord's constant hazard is a claim at this level: for a given agent and a given kind of work, each additional unit of human time multiplies success by the same factor. Mittal (2026), Khanal et al. (2026), Potash (2026) and the synthetic-horizon benchmarks measure here.

**Level C: the suite.** Pool many task families of different kinds, rate each task by the time a human takes, and fit success against that rating. This is what Kwa et al. (2025) measure and what the 50 percent horizon summarises. At this level a task's length is both a

duration and a proxy for its difficulty, and the two cannot be separated, because short tasks and long tasks are different tasks.

The accounts are claims at different levels. The logistic fit is a level-C description. The constant-hazard account is a level-B hypothesis that Ord tested against level-C data. The self-conditioning result is a level-A measurement whose authors draw a level-B implication. None of the three is wrong about the level at which it was made. The question for forecasting is how the levels connect.

### 4.2 From the run to the family: steps per unit time

Level A counts steps; level B counts human minutes. Connecting them requires a rate - how many agent steps a minute of human work corresponds to - and that rate is neither constant nor reported. In Sinha et al.'s task the plan is supplied and each turn is a retrieval and an addition; an hour of human work in an HCAST task may require exploration, debugging and decisions about what to do next (Rein et al., 2025). Kwa et al. scored their tasks on sixteen "messiness" factors, such as a novel situation, a finite resource or a dynamic environment; the mean score was 3.2 of 16 and none exceeded 8, against an estimated 9 to 15 for a task like writing a good research paper. Holding length fixed, success fell by roughly 8.1 percent per point of messiness (Kwa et al., 2025). The steps that fail in messy work are therefore not the same kind of step as the ones Sinha et al. measured, and the exponents do not transfer.

There is a second, subtler mismatch. A within-run hazard concerns failure from which the run does not recover. If an agent can notice and repair errors, a step error is not a task failure, and the per-minute hazard at level B reflects the rate of unrecovered errors only. The earlier record in this journal documented that recovery is regime-dependent - common where a task is easy for the model, rare where it is hard (Mahendrakar, 2026). Kwa et al.'s own qualitative analysis points in the same direction over time: among 31 failed GPT-4 runs, 12 were classified as repeating failed actions; among 32 failed o1 runs, 2 were (Kwa et al., 2025). They attribute part of the trend to "greater reliability and ability to adapt to mistakes" (Kwa et al., 2025, v1). If the trend in the horizon is partly a trend in recovery, the level-A hazard that matters is changing in form, not only in size,

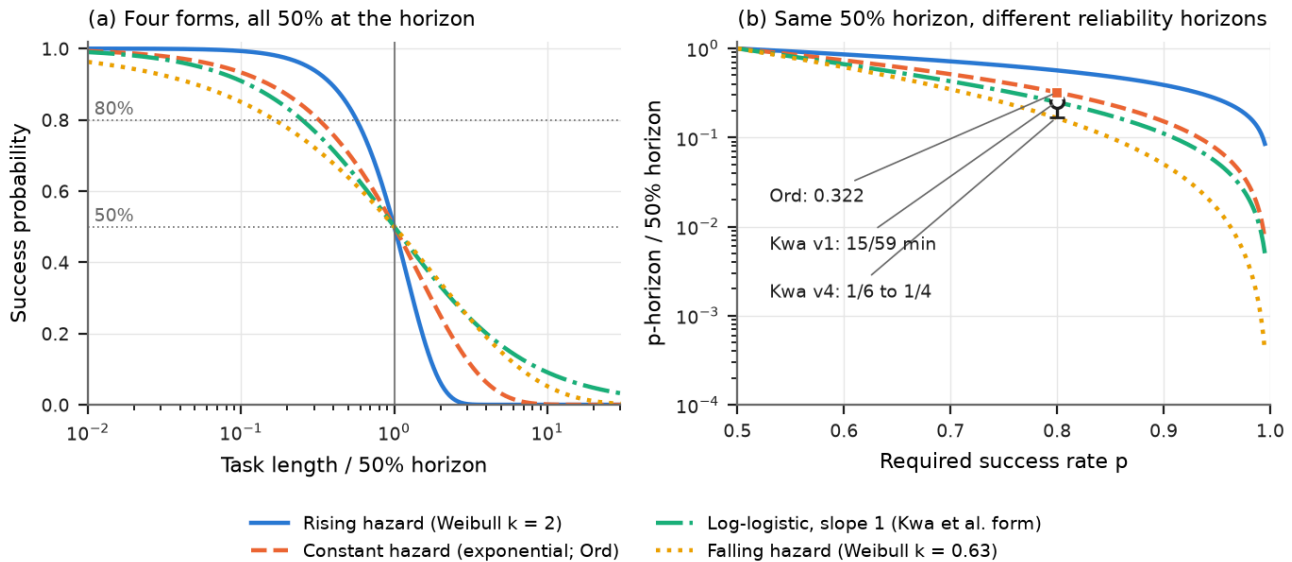


**Figure 1.** Four failure shapes calibrated to the same 50 percent horizon. (a) Success probability against task length as a multiple of the 50 percent horizon, for a rising hazard (Weibull shape 2), a constant hazard (exponential, Ord 2025), the log-logistic form with slope 1 used by Kwa et al. (2025), and a falling hazard (Weibull shape 0.63). (b) The p-reliability horizon as a fraction of the 50 percent horizon for the same forms. The curves are derived from the functional forms and are not data. The marked points are published: Ord's constant-hazard prediction of 0.322 at 80 percent; Kwa et al. (2025, v1), Claude 3.7 Sonnet, 80 percent horizon of about 15 minutes against 59 minutes (0.254); Kwa et al. (2025, v4), 80 percent horizons 4 to 6 times shorter than 50 percent horizons (bar from 1/6 to 1/4).

from one model generation to the next.

#### 4.3 From the family to the suite: unobserved heterogeneity

The connection between levels B and C is a problem with a long history in reliability engineering, demography and econometrics, and its main result is unwelcome for anyone hoping to read a failure process off a pooled curve.

Proschan (1963) gave the theoretical explanation of a decreasing failure rate observed in pooled equipment data: pooling units that each fail at a constant but different rate produces a mixture whose hazard falls with time, because the units with high rates fail early and leave a survivor population increasingly made of low-rate units. Vaupel et al. (1979) examined the corresponding effect of heterogeneity in individual frailty on the dynamics of mortality. In econometrics the problem became one of identification. Observed duration dependence in pooled data can be true, a property of each unit's hazard, or spurious, a product of heterogeneity across units, and Elbers and Ridder (1982) and Heckman and Singer (1984) studied what must be assumed or observed for the two to be told apart. A single pooled curve of success against length, with no information about which unit is which, is the case in which they cannot be. Ord's earlier work on the Lindy effect reaches a related conclusion from a different direction: a population can show long-tailed survival "even in cases with constant (or increasing) hazard rates - so long as there is a probability distribution over the size of that rate" (Ord, 2023).

Applied to METR's suite, the consequence is direct. Tasks of the same human length differ widely in difficulty for agents; that is why Kwa et al. trust the slope of their trend more than any one model's horizon. Success on tasks is also strongly correlated across models, with an average correlation of about 0.73 (Kwa et al., 2025), which says that some tasks are persistently harder than their human time suggests. Task-specific difficulty per unit time is exactly the kind of heterogeneity that makes a pooled hazard fall. So the published 80 percent ratios of one quarter to one sixth, which imply a falling hazard if read as a single survival curve, are equally well explained by constant hazards within each task family with different rates across families, and, with enough heterogeneity, by rising hazards within each family. The pooled curve does not distinguish these.

#### 4.4 A dataset that reports both levels

A 2026 preprint reports measurements at levels A and B in the same data, and it shows how easily a within-run figure can be misread. Mittal (2026) ran nine models on four synthetic task families at five horizons, 10,664 analysed trajectories in all. Comparing geometric, threshold and linear decay shapes by AIC across horizons, the geometric form, the step-count analogue of Ord's constant hazard, won in 28 of the 36 model-task cells; the remainder followed a threshold or cliff shape. Within long trajectories, Mittal reports that mean per-step accuracy, pooled across trajectories, fell from 0.58 in the first third to 0.44 in the last third, and reads this as the hazard "rising instead" (Mittal, 2026). The same paper reports that once an agent errs it rarely recovers, so that "errors behave as absorbing states" (Mittal, 2026).

Under absorbing errors, pooled per-step accuracy includes the steps of runs that have already failed, and those runs make up a growing share of the later steps; it would therefore fall with position even if the hazard of the first error were exactly constant. The 0.58 to 0.44 decline is Mittal's measurement and Mittal's reading, but it does not by itself show a rising hazard, and a geometric law across lengths sits beside it without tension.

The distinction holds beyond this one study. A geometric law across lengths is a statement about the probability of completing  $H$  steps for each  $H$ ; it is fitted to end-of-run outcomes. A within-run hazard is a statement about the conditional probability of an unrecovered error at the next step, given that the run has not yet failed. Pooled per-step accuracy is neither: it mixes the hazard with the share of runs that have already failed, and pooling over task instances of different difficulty adds a further layer. Mittal's study is a single-author preprint, and its pooled within-trajectory figure mixes models and families. The structural point does not depend on it: a good fit of one functional form across lengths is not evidence about the hazard inside a run, and the within-run hazard has to be measured conditionally on survival before it can be read at all.

The same lesson appears in reverse in Khanal et al. (2026), who ran ten open models on 396 tasks in four human-duration buckets and three domains. Mean  $\text{pass@1}$  fell from 76.3 percent on short tasks to 59.8 percent on medium ones, faster than the authors' geometric baseline predicts, which they read as positive correlation between step errors. It then fell to 50.5 percent on long tasks and rose slightly, to 52.1 percent, on very long ones (Khanal et al., 2026). A steeper-than-geometric drop followed by a plateau is what a population of task families with very different rates would produce; Khanal et al. also report that the decline was strongly domain-stratified, with a partial-credit score falling from 0.90 to 0.44 in software engineering and barely moving, from 0.74 to 0.71, in document processing. The domain pattern is heterogeneity made visible. A single hazard shape fitted to the pooled buckets would be fitting the mixture.

#### 4.5 Time as difficulty, time as exposure

Section 3.1 noted that the logistic fit and its IRT descendants treat human time as a difficulty coordinate, while Ord treats it as exposure to a constant risk. The distinction matters for what an extrapolation claims. If time is a difficulty coordinate, then a month-long task is a task whose difficulty sits at  $\log(\text{month})$  on the scale, and the forecast that agents will reach a 50 percent chance on such tasks extrapolates the rise in ability along an axis that has been calibrated only up to roughly the longest tasks in the suite. If time is exposure, a month-long task is a long sequence of shorter ones, and the forecast is a statement about the product of survival

over that sequence. The two readings agree at the midpoint they are fitted to and diverge far from it, which is precisely where a one-month forecast lives. Liu et al. (2026) find item difficulty linear in log human time over the range they observe; nothing in their analysis or in Kwa et al.'s establishes that the linearity continues to tasks several times longer than any in the calibration set, and Kwa et al. state that measuring long horizons requires long human baselines that are "impractical to construct" (Kwa et al., 2025).

## 5. What the evidence shows at each level

### 5.1 Level A: rising, flat and cliff-shaped hazards all appear

The within-run evidence does not support one hazard shape. It supports at least three, in different conditions.

**Rising.** Sinha et al. (2025) report per-turn accuracy falling over 100 turns for non-thinking models, with an identifiable self-conditioning component. Laban et al. (2025) found that models given an instruction split across turns lost 39 percent of their performance on average, decomposed into a 16 percent loss of aptitude and a 112 percent increase in unreliability, and summarised the mechanism as getting lost after a wrong turn and not recovering. Zhang et al. (2023) showed that models over-commit to early mistakes and then produce false claims they can separately recognise as false, which they called hallucination snowballing. Pereira and Zuidema (2026) found that two frontier reasoning models encode a near-perfect representation of the Tower of Hanoi state space at the start of planning, fail most problems of the flat-to-flat variant above three rings, and trace the failure to a representation that decays during planning. Each of these is a mechanism by which the hazard rises with the length of a run. Mittal (2026) reports pooled per-step accuracy falling from 0.58 to 0.44 between the first and last thirds of long trajectories and reads it as a rising hazard; given the absorbing errors the same paper reports, a constant first-error hazard would produce the same fall (Section 4.4), so this figure does not by itself discriminate.

A rising hazard need not come from errors in context; it can come from context length itself. Even with perfect retrieval of the relevant information, performance fell by 13.9 to 85 percent as input length grew within models' claimed limits (Du et al., 2025), and models use information in the middle of long contexts less well than information at either end (Liu et al., 2023). The two mechanisms are not always separable, and the evidence on which dominates in agent loops conflicts. Sinha et al. (2025) separate them with healed histories and find both. Mittal (2026) finds that compressing the context made the decline steeper, not shallower, which counts against a long-context explanation. In Vending-Bench, runs that derailed into "meltdown" loops did so with no clear correlation to the point at which the context

window filled (Backlund and Petersson, 2025).

**Flat.** Potash (2026) measured six frontier models on a two-agent game in which a questioner must identify one of  $N$  documents with exactly  $\log_2(N)$  yes-or-no questions. Win rate decayed geometrically with the number of rounds, at about 0.93 per round, and per-round failure rates were flat across the horizon: errors compounded because more rounds had to succeed, not because rounds grew harder. In Sinha et al. (2025), Qwen3 thinking models showed stable turn-100 accuracy whatever the error rate injected into their history. In these settings, the constant hazard holds where it can be checked.

**Cliff.** In Mittal (2026), eight of 36 model-task cells were better described by a threshold than by geometric decay: strong models held near-perfect accuracy up to a critical horizon and then failed. Shojaei et al. (2025) reported a complete accuracy collapse of reasoning models on puzzles beyond a complexity threshold. Whether that cliff is a property of the models is disputed: Lawsen (2025) argued that some reported failures exceeded output token limits and that some River Crossing instances were unsolvable; Dellibarda Varela et al. (2025) replicated the Tower of Hanoi experiments with stepwise prompting and found that failures were not purely output constraints, with models still stumbling at around eight disks; and Khan et al. (2025) reframed the collapse as an "agentic gap" closed when the model may use tools. The disagreement is itself informative here: whether the hazard has a cliff depends on what the model is allowed to do and how its output is scored.

Nothing in this evidence licenses a single within-run hazard shape for agents in general. It licenses something narrower: that the shape is a property of a model, a harness and a task, and that it ranges from flat to sharply rising to cliff-like across combinations that are all in current use.

### **5.2 Level B: geometric fits across lengths are common and uninformative about level A**

At level B the evidence is more uniform. Mittal (2026) finds geometric decay across horizons in 28 of 36 cells, and on its one genuinely agentic tool-use family, seven of nine models (per the source's Table 2) were at or near ceiling at two steps, and none of those seven kept more than a third of that success by sixteen steps, a steep decay across lengths. Potash (2026) finds a clean geometric law with a per-round factor of about 0.93. Tran-Truong and Le (2026) propose fitting absorbing discrete-time Markov chains to agent traces. Their held-out test is a controlled one: for seven MAST-style frameworks they generated traces from a known absorbing chain with noisy features, and the fitted chains reproduced held-out reliability decay curves to within about 0.05 and passed a two-sample Kolmogorov-Smirnov test on all seven. That shows the fitting pipeline recovering a chain from Markov-generated data; the

authors state that it "does not prove that operational agent traces are Markov" (Tran-Truong and Le, 2026). Bogdanov et al. (2026), in a different kind of task - a 20-step game in which an agent repeatedly chooses whether to keep delaying a reward - use Kaplan-Meier curves and discrete-time hazard regression and report distinct temporal regimes, with early failures more impulsive and later ones framed in terms of fatigue and cost-benefit reasoning. Because the hazard runs over the steps of a single episode, it is strictly a within-run quantity, level A in the terms of Section 4.1, and it is noted here for its method: a hazard that changes in character over time, measured with the standard survival toolkit (Kaplan and Meier, 1958; Cox, 1972).

The lesson of Section 4.4 applies throughout: a good level-B fit says that the end-to-end success probability is well described by a given form over the lengths sampled. It does not reveal whether the per-step process is memoryless.

### **5.3 Level C: a falling pooled hazard, a thicker human tail, and domain spread**

At level C the published 80 percent ratios imply a falling pooled hazard (Section 3.4). Three further observations bear on why.

First, humans. Ord (2025) noticed that the human success curve in Kwa et al.'s data declined more slowly than a constant hazard predicts: humans succeeded about half the time at the 1.5-hour mark, which would imply under 1 percent (about 0.4 percent, since 12 hours is eight half-lives) at 12 hours under a constant hazard, but were still above 20 percent there. Ord gives "6.25% at 12 hours", a sequence that halves once per doubling of task length rather than once per 1.5 hours; the correction strengthens his point. He offered two explanations - that humans are better at repairing earlier failures, or that the curve pools humans of different ability - and did not choose between them. RE-Bench, one of the three components of METR's suite, reports a related pattern from another angle: the best agents scored four times higher than human experts with a two-hour budget, but humans showed better returns to additional time, narrowly exceeding the top agent at eight hours and doubling its score at 32 hours (Wijk et al., 2024). Better returns to time is what a falling hazard, or a recovery process, produces.

Second, domains. Kwa et al. (2025) expect time horizons to differ in other domains, and found a much shorter doubling time on SWE-bench Verified, around 70 days against 143 days on their own suite for 2024 models, which they attribute partly to annotators' time estimates for easy tasks being too short. Khanal et al. (2026) found strong domain stratification. Wang et al. (2026), with more than 3,100 trajectories across four agentic domains, report that the horizon at which agents break down differs by domain and that the composition of failures shifts as horizons grow, with planning and

memory failures becoming dominant. Each of these is evidence of heterogeneity in the rate, the condition under which a pooled curve's shape says least about its components.

Third, the shape at the suite level may be partly an artefact of how difficulty and ability are scaled. Xing et al. (2026) show that the apparent shift of gains towards harder tasks on METR's data is largely reproduced by a single Rasch model with rising ability, so that the pooled curve's shape over time is largely a ceiling effect of a fixed response curve; a smaller hard-task effect survived their control only on a benchmark where scaffold changes could be ruled out. They draw the comparison with the observation that apparently emergent abilities can be products of the metric (Schaeffer et al., 2023). The same caution applies to reading a failure process off a pooled time-horizon curve: its shape is a joint property of agents, tasks and the scale on which difficulty is placed.

#### **5.4 Table 1: the measured values the argument rests on**

Table 1 consolidates the published values used in Sections 3 to 7. Each row names the source that reported its value.

### **6. The hazard is a property of a system, and systems are being redesigned around it**

If the within-run hazard were a fixed property of a model, the dispute between accounts would be a dispute about nature. The evidence suggests instead that it is a property of the model together with its harness, and that harnesses are being engineered specifically to change it.

The clearest case is Meyerson et al. (2025). Starting from the observation that Tower of Hanoi solutions derail after at most a few hundred steps, they decomposed the task into single-move subtasks, sampled several independent responses for each, accepted a move only when one answer led the others by a margin, and discarded responses showing signs of confusion, such as malformed output. With a small non-reasoning model this solved the 20-disk problem, over one million steps, with no errors. Their own analysis assumed independent errors across steps and noted a few steps with unexplained, unusually high error rates, which resampling and filtering were sufficient to overcome (Meyerson et al., 2025). The system's effective per-step hazard was driven close to zero and held flat not by a better model but by a different process around it.

Pai and Xian (2026) report a smaller but similar case. A mixture-of-experts model with about 5.5 billion active parameters carried the state of an MD5 computation across 196 dependent tool calls, and more than half of the runs that finished produced the right hash. In the harder setting, with every arithmetic tool replaced by a second model, the authors identify two conditions for

success, neither a change to the weights: keeping the model's own reasoning in its context at every turn, since stripping it made success collapse, and majority voting over a thinking-enabled worker to remove arithmetic slips. Set beside Sinha et al. (2025), this is instructive. In one study, removing prior turns from context mitigated self-conditioning; in the other, keeping the model's own reasoning in context was essential. What a harness keeps in context changes the within-run hazard, and in either direction depending on what is kept.

Three further results point the same way. Sinha et al. (2025) find that thinking removes self-conditioning in the models they tested. Zhu et al. (2025) build a debugging framework for agent trajectories that isolates a root-cause error and feeds back a correction, and report up to 26 percent relative improvement in task success across three benchmarks, which is a change in the recovery term of the hazard. And the shift in Kwa et al.'s failure categories between GPT-4 and o1, from repeated failed actions towards premature abandonment, says that the dominant failure mode changed between generations: a newer model that stops too early fails by a different process from an older one that loops (Kwa et al., 2025). Wang et al. (2026) report a similar shift in failure composition with horizon length. In survival terms these are competing risks, and the mix of risks is itself moving.

For extrapolation, this has an uncomfortable implication. The 50 percent trend pools models whose within-run hazards differ in form - non-thinking and thinking, with older and newer harnesses. Xing et al. (2026) make the corresponding point for difficulty: because each model generation tends to arrive with its own updated scaffold, an improvement on agentic benchmarks cannot be credited to either one alone. A forecast that extrapolates the trend is implicitly extrapolating the rate at which harness designers and model trainers will continue to reshape the hazard. That is not an argument against the forecast. It is an argument that the forecast is about the joint progress of models and the systems built around them, and that no failure model fitted to one generation describes the next.

### **7. Which forecasts survive not knowing the hazard**

#### **7.1 The 50 percent trend is nearly indifferent to the dispute**

Every functional form considered in this paper is calibrated to pass through one half at the agent's measured 50 percent horizon. The horizon itself does depend a little on the family fitted - Ord's exponentials and Kwa et al.'s log-logistics are fitted to the same binned data and need not cross one half at exactly the same length - but the doubling time is estimated from the trend in midpoints across models, and Kwa et al. (2025) report that the fitted trend is not sensitive to

**Table 1.** Published measurements bearing on the failure models behind horizon extrapolation, with the source of each row. Levels refer to Section 4.1 (A: within a run; B: across lengths within a task family; C: across a pooled task suite).

| Source                 | Level         | Setting                         | Reported value                                                                    |
|------------------------|---------------|---------------------------------|-----------------------------------------------------------------------------------|
| Kwa et al. (2025), v1  | C             | METR suite, 2019-2025 models    | 50% horizon doubling 212 days (95% CI 171-249); 80% doubling 213 days             |
| Kwa et al. (2025), v1  | C             | Claude 3.7 Sonnet               | 50% horizon 59 min; 80% horizon about 15 min (ratio 0.25)                         |
| Kwa et al. (2025), v4  | C             | METR suite                      | 50% doubling 207 days (166-240); 80% doubling 204 days; 80% horizons 4-6x shorter |
| Kwa et al. (2025), v4  | C             | SWE-bench Verified              | doubling about 70 days, vs 143 days on own suite for 2024 models                  |
| Kwa et al. (2025), v4  | C             | Failed runs, GPT-4 1106 vs o1   | repeating failed actions 12/31 vs 2/32; premature abandonment 8/31 vs 16/32       |
| Kwa et al. (2025), v4  | C             | HCAST tasks                     | about 8.1% lower success per messiness point, length held fixed                   |
| Ord (2025)             | B (on C data) | Constant-hazard derivation      | h80/h50 = 0.322; h90 about 1/7; h99 about 1/70                                    |
| Sinha et al. (2025)    | A             | Supplied-plan execution         | Qwen3-32B below 50% task accuracy within 15 turns                                 |
| Sinha et al. (2025)    | A             | Single-turn execution           | GPT-5 2,176 steps; Claude-4 Sonnet 432; Grok 4 384; Gemini 2.5 Pro 120            |
| Sinha et al. (2025)    | A             | Qwen3 thinking models           | turn-100 accuracy stable under injected history errors                            |
| Mittal (2026)          | B             | 9 models, 4 synthetic families  | geometric decay preferred by AIC in 28 of 36 cells                                |
| Mittal (2026)          | A             | Long trajectories, pooled       | per-step accuracy 0.58 (first third) to 0.44 (last third)                         |
| Khanal et al. (2026)   | B/C           | 10 models, 396 tasks            | mean pass@1 76.3 / 59.8 / 50.5 / 52.1% by duration bucket                         |
| Potash (2026)          | A/B           | log2(N)-questions game          | per-round success about 0.93, failure rate flat across rounds                     |
| Laban et al. (2025)    | A             | Instructions sharded over turns | 39% average drop; aptitude -16%, unreliability +112%                              |
| Meyerson et al. (2025) | A             | Towers of Hanoi, 20 disks       | over one million steps, zero errors, with decomposition and voting                |
| Costa (2026)           | C             | METR Time Horizon 1.0 to 1.1    | log-scale linking slope 1.206 (95% CI 1.021-1.390), 7 systems                     |
| Whitfill et al. (2025) | C             | Compute-slowdown model          | 1-month 80% horizons 7 years later than simple trend, in one projected case       |

regularisation, to how tasks are weighted, or to the choice between weighted and ordinary least squares. The failure-model dispute is a dispute about the shape of each agent's curve away from its midpoint. It barely touches the trend in midpoints.

That is reassuring only about one source of doubt. The 50 percent trend has its own, well documented problems, and they are larger than anything the failure model contributes. The instrument changes: across a revision of METR's own task suite, seven systems measured on both versions show a log-scale linking slope of 1.206, with a 95 percent interval of 1.021 to 1.390, and 52 of the 71 substantive quantitative events in one forecasting audit come from that one programme (Costa, 2026). The domain matters: the same method gives a doubling time of about 70 days on SWE-bench Verified (Kwa et al., 2025). Single runs are noisy: across 60,000 SWE-bench Verified trajectories, choosing a different single run moved a model's pass@1 by between 2.2 and 6.0 percentage points (Bjarnason et al., 2026). The trend may not continue: Kwa et al. fit an exponential because it fits well over twelve frontier models and decline to fit saturating or super-exponential forms they cannot distinguish from it, and Whitfill et al. (2025) show how slower compute growth would slow it. And the task suite may not stand for real work: an RCT of experienced open-source developers found that early-2025 AI tools increased completion time by 19 percent, against forecasts of large speed-ups (Becker et al., 2025), and the

most competitive agent on a simulated company's workplace tasks completed 30 percent of them (Xu et al., 2024).

The general forecasting literature offers a pattern that fits the position of the 50 percent trend. Aggregate benchmark performance is markedly more predictable than performance on individual tasks: extrapolating across an order of magnitude of compute, average errors were 6 percentage points for an aggregate and 18 for individual tasks (Owen, 2024). Low-dimensional capability measures predict agent performance from non-agentic benchmarks (Ruan et al., 2024), forecasting agent benchmark scores from release date through an intermediate capability measure validates in backtests (Pimpale et al., 2025), and statistical frameworks that place models and benchmarks on one scale can track progress without assuming how capabilities evolve (Ho et al., 2025). The 50 percent horizon is a summary of this aggregate kind: a midpoint of a curve fitted over many tasks. Its predictability is not evidence that the curve's tails are predictable.

## 7.2 Every reliability horizon above one half is assumed, not measured

Kwa et al. (2025) are explicit that their data cannot measure horizons at very high success rates: doing so would need very large task sets with near-zero label noise covering the population of tasks they stand for. What they report at 80 percent is measured, and its

doubling time, 204 days in the current version, is close to the 50 percent doubling time. That is evidence that, over the period observed, the ratio of the two horizons stayed roughly fixed. It is not evidence about where the 95 or 99 percent horizons are.

Table 2 shows what the choice of shape does to those horizons. If the 50 percent horizon keeps doubling every seven months and the ratio of each reliability horizon to it stays fixed, the 99 percent horizon reaches any given length a fixed number of doublings later. Under a rising hazard that lag is about 3.1 doublings, roughly 21 months; under Ord's constant hazard it is 6.1 doublings, roughly 43 months, which matches his estimate of about four years; under the falling-hazard shapes consistent with the published 80 percent ratios it runs from 6.6 doublings, about 46 months, to 9.7, about 68 months. These lags are derivations from functional forms, not forecasts. They say that the interval between the date at which agents first complete a month of human work half the time and the date at which they do so 99 times in a hundred is, on the evidence reviewed here, somewhere between under two years and nearly six. The evidence cannot narrow that range, because it is set by the shape of the curve in a region where nothing has been measured.

There is a further ambiguity in what a level-C reliability horizon means. Kwa et al. run agents several times per task, so their data contain per-task success rates, but the fit pools runs across tasks. "Eighty percent success on tasks of length  $t$ " can mean that every task of that length succeeds 80 percent of the time, or that 80 percent of such tasks always succeed and the rest never do, or anything between. A deployment cares about a particular task done repeatedly, which is the quantity  $\tau$ -bench's  $\text{pass}^k$  metric targets: on that benchmark, the best function-calling agents of mid-2024 succeeded on fewer than half of tasks and, in the retail domain, succeeded on all eight of eight repeated trials of a task less than a quarter of the time (Yao et al., 2024). The suite-level 80 percent horizon does not say which of the readings holds, and they imply very different things for a user who must rely on one task being done reliably. This is heterogeneity again, now at the level of tasks rather than task families.

### ***7.3 Tasks far beyond the horizon are where the forms disagree most***

Panel (a) of Figure 1 shows the second place where the forms diverge: tasks much longer than the 50 percent horizon. At ten times the horizon, the derived success probability is about 0.001 under a constant hazard, about 0.01 under a Weibull with shape 0.82, about 0.09 under the log-logistic with slope one, and effectively zero under a rising hazard of shape two. The question "could this agent complete a task ten times longer than its horizon, given many attempts?" - which matters for both usefulness and risk - has answers spanning orders of

magnitude depending on which shape is assumed. The log-logistic's thick tail is a property of the fitted family, and at level C it is partly a property of the task mix; neither is evidence that any single long task can be completed with that probability.

The month-long forecast in Kwa et al. (2025) is a forecast about the date at which the 50 percent horizon reaches a month, a quantity the failure model barely affects; on the central estimate it falls in 2029, and the authors note that the 2024-2025 trend would bring it sooner. What the failure model affects is everything said afterwards about such a system: how often it would succeed at a task of a given length, how much more often with retries, and how long the wait would be until it succeeds reliably.

## **8. A reading procedure for horizon claims**

Algorithm 1 states the decision structure this analysis implies for reading, or making, a claim about agent task-length horizons. It is a procedure for classifying a claim by what it assumes, not a statistical method, and it is the author's synthesis.

The procedure's first branch is deliberately permissive. Most of what the 50 percent horizon is used for - comparing models, tracking the pace of progress on a fixed suite - passes it, subject to the validity checks that its authors and critics already discuss. The second branch is where most claims about deployment, safety margins and reliability land, and its usual outcome is that the claim is resting on an assumption about failure shape that the data do not test.

## **9. Objections**

**Ord already raised heterogeneity, so the point is not new.** Ord (2025) did raise it, in a short paragraph, as a reason why the constant-hazard mechanism might hold for every task while being invisible in the aggregate. This paper builds on that observation rather than claiming it. What is added is the separation of three levels at which the published accounts were made, the sorting of the 2025-2026 evidence by level, and the derivation of which forecasts depend on the part that heterogeneity leaves unidentified. Ord's own reading of the 80 percent ratio as "within the margin of error" of the constant-hazard value was made against the first version of Kwa et al.; the current version's range of one sixth to one quarter lies further from it.

**Nobody forecasts 99 percent reliability from these data.** Kwa et al. avoid doing so. Others do not: Ord (2025) derives a reliability ladder and the lag to reach each rung; Whitfill et al. (2025) project when one-month horizons will be reached at 80 percent reliability; and the deployment literature argues that production reliability, not 50 percent success, is what matters (Mittal, 2026; Khanal et al., 2026). The 80 percent horizon is reported and trended by the metric's authors.

**Algorithm 1.** Reading a task-length horizon claim against the failure model it assumes.

```

Input: a claim Q about agent success at task length t and reliability p
1: Identify the level of Q:
   A = within a run (steps); B = one task family across lengths;
   C = a pooled suite rated by human time
2: Identify the role of time in Q: difficulty coordinate, or exposure to a risk
3: if p = 0.5 and t is within the measured range of task lengths:
4:   Q depends on the trend in midpoints only; failure shape is irrelevant
5:   check instrument version, domain, run-to-run noise, external validity
6:   return "shape-free; judge on trend and validity"
7: if p > 0.5 or t lies well beyond the longest measured tasks:
8:   Q depends on the curve away from its midpoint
9:   if the level of Q is C (suite) and the claim is about one task or user:
10:    flag heterogeneity: suite reliability != per-task reliability
11:   if the shape is taken from a fitted family (exponential, log-logistic):
12:    report Q under at least two families calibrated to the same h50, h80
13:   if the answers differ by more than the claim's stated precision:
14:    return "shape-dependent; state the assumed hazard"
15:   if within-run hazard evidence exists for this model and harness:
16:    use it only for level-A or level-B claims about the same task family
17:   if the harness differs from the one measured (memory, voting, thinking):
18:    treat the measured shape as not transferable
19: return "assumption-bearing; the reliability horizon is not measured"

```

Any statement at 80 percent beyond the measured range, or above it at any length, uses a shape.

**Thinking fixes self-conditioning, so the rising hazard is a solved problem.** The fix was shown for Qwen3 thinking models on a synthetic execution task (Sinha et al., 2025). Mittal (2026) reports per-step accuracy falling within long trajectories, pooled across open and proprietary models and across task families, and reads it as a rising hazard, subject to the caveat of Section 4.4; and multi-turn degradation was found across all fifteen models tested by Laban et al. (2025). More importantly for this paper's argument, a fix that changes the shape of the hazard between model generations is a reason not to extrapolate any one generation's shape. Whether the problem is solved and whether a fitted shape transfers are different questions, and the second has a negative answer either way.

**If time is a difficulty coordinate, survival analysis is the wrong frame.** If the IRT reading is right, the reliability ladder has no mechanistic basis at all, and the 80 and 99 percent horizons are whatever the chosen link function says they are at those quantiles. That makes the case of Section 7 stronger, not weaker: the tail of a logistic link fitted around its midpoint is no better identified than the tail of a survival curve.

**Short-task benchmarks understate progress, so worrying about tails is a distraction.** Sinha et al. (2025) argue that small gains in per-step accuracy compound into large gains in horizon, so that apparent diminishing returns on short tasks are an illusion. Their argument is about the middle of the curve and is compatible with everything here. The same compounding arithmetic, applied to the tail, is what makes reliability forecasts sensitive to small errors in the assumed hazard.

## 10. What would settle the open part

Six measurements would resolve most of what this paper leaves open. None requires new benchmarks; several use data that evaluation programmes already collect.

**M1. Within-run time to irrecoverable failure on realistic tasks.** For agent runs on suite tasks, annotate the point after which the run could not recover, and estimate the within-run hazard by Kaplan-Meier methods per task family (Kaplan and Meier, 1958). This is the level-A quantity that the self-conditioning debate is about, measured on the tasks the horizon is computed from rather than on synthetic ones. Trajectory-annotation datasets that localise root-cause failures already exist for other benchmarks (Zhu et al., 2025; Wang et al., 2026).

**M2. Per-task success distributions at each length.** Evaluation programmes that run several attempts per task should report the distribution of per-task success rates by length bin, not only the pooled rate. Mixed-effects survival models with task-level frailty (Cox, 1972; Vaupel et al., 1979) would then separate the within-task shape from the across-task spread, the separation whose identification Elbers and Ridder (1982) and Heckman and Singer (1984) studied, and which a pooled curve cannot provide.

**M3. Length-controlled versions of realistic tasks.** The cleanest level-B evidence comes from synthetic tasks whose length can be dialled (Sinha et al., 2025; Mittal, 2026; Pai and Xian, 2026). Realistic task families with a length parameter - the same refactoring across more files, the same analysis over more data - would test whether a constant hazard per unit of human time holds where it is used.

**M4. Family-sensitivity bands on every reliability horizon.** Any reported 80, 90 or 95 percent horizon should be accompanied by the same quantity under at least two functional forms calibrated to the same data, as

Ord (2025) asked for. Where they disagree, the disagreement is the honest error bar on the assumption.

**M5. Harness ablations of the hazard shape.** For one model, measure the within-run hazard with and without history pruning, voting, reasoning in context and thinking, on the same task family. Section 6 suggests these change the shape; a direct measurement would say by how much, and would show whether a horizon trend that pools harnesses is pooling shapes.

**M6. The human comparison, decomposed.** Ord (2025) noticed a thicker tail in human success and offered recovery and heterogeneity as rival explanations; RE-Bench found better human returns to time (Wijk et al., 2024). Applying M1 and M2 to the human baselines METR already holds would settle which, and would say whether the human-agent gap is a gap in the rate of failure or in its shape.

The prediction that would most change the picture is simple to state. If M1 and M2 found constant within-task hazards with a wide spread across tasks, Ord's mechanism would hold at the level he proposed it and the reliability ladder would be computable per task family, though not for the suite. If they found rising within-task hazards that harnesses flatten, reliability horizons would depend on the harness more than on the model, and would need to be measured per deployment.

## 11. What this paper does not show

The paper builds its argument from published measurements, so every claim about agent behaviour is a claim about what cited sources report. Four limitations deserve stating flatly.

First, it does not show that any of the three accounts is wrong. Each is accurate at the level at which it was made. What it shows is that the levels have been conflated, and what follows for forecasting when they are separated.

Second, the derivations in Table 2 and Figure 1 depend on the families chosen. Weibull and log-logistic forms are standard and span rising, constant and falling hazards, but other two-parameter families calibrated to the same two quantiles would place the 99 percent horizon elsewhere. That strengthens rather than weakens the claim that the tail is unidentified, but the specific factors quoted - about eight across falling shapes, about 1.8 between two forms matched at two points - are properties of these families, not of agents.

Third, several of the most relevant measurements come from recent preprints that have not been peer reviewed, and three are single-author studies (Mittal, 2026; Potash, 2026; Costa, 2026). The argument from heterogeneity does not depend on any of them, since it follows from Proschan (1963) and Vaupel et al. (1979).

Fourth, the self-conditioning evidence that drives the rising-hazard account comes mainly from one synthetic

task with supplied plans (Sinha et al., 2025). How far its exponents transfer to work in which planning, exploration and judgement dominate is not known, and the mapping from agent steps to human minutes that would connect level A to level B has not been measured by anyone. Kwa et al.'s finding that messier tasks are harder at equal length, and Wang et al.'s finding that the composition of failures shifts with horizon, suggest that it does not transfer simply.

## 12. Conclusion

The time-horizon metric condenses agent capability into the human task length at which an agent succeeds half the time, and three accounts of why success falls with length have grown up around it. They are usually read as competitors. Read carefully, they are claims about three levels of the same data: the hazard inside a run, the hazard per unit of a task family's length, and the success curve across a pooled suite. The levels are joined by the aggregation that survival analysis calls unobserved heterogeneity, under which a falling pooled hazard is compatible with constant or rising hazards inside every task. The within-run evidence itself is split - rising, flat and cliff-shaped hazards all appear - and harnesses built around voting, decomposition, context management and thinking change the shape between generations.

Two conclusions follow. The 50 percent trend is nearly indifferent to which account is right, because all are calibrated to it; its weaknesses lie elsewhere, in instruments, domains, noise and external validity. Every statement above 50 percent reliability, or about tasks far beyond the measured horizon, inherits the shape that the data do not identify. Holding the 50 percent horizon fixed, curves consistent with the published 80 percent ratios disagree about the 99 percent horizon by a factor of about eight, and the delay between half-time and near-certain success at any given task length ranges, across the shapes reviewed, from under two years to nearly six. Those reliability horizons are not measured. They are assumed, and the assumption is the one in dispute.

## References

- Backlund, A., & Petersson, L. (2025). Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents. arXiv:2502.15840. doi:10.48550/arXiv.2502.15840.
- Becker, J., Rush, N., Barnes, E., & Rein, D. (2025). Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. arXiv:2507.09089. doi:10.48550/arXiv.2507.09089.
- Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. arXiv:1506.03099. doi:10.48550/arXiv.1506.03099.
- Bennett, S. (1983). Log-Logistic Regression Models for Survival Data. *Applied Statistics*, 32(2), 165-171. doi:10.2307/2347295.
- Bjarnason, B. H., Silva, A., & Monperrus, M. (2026). On Randomness in Agentic Evals. arXiv:2602.07150. doi:10.48550/arXiv.2602.07150.
- Bogdanov, I., Manakina, O., & Lung, C.-H. (2026). Evaluation of Multi-Turn Consistency in LLM Agents: Survival Analysis and

- Failure-Rationale Taxonomy. ICLR 2026 Workshop on Logical Reasoning of Large Language Models. arXiv:2609.29508. doi:10.48550/arXiv.2609.29508.
- Chen, M., Wang, L., & Qu, B. (2026). The Horizon Gap: Planning, Memory, Execution, Training, and Evaluation for Long-Horizon LLM Agents. arXiv:2608.06663. doi:10.48550/arXiv.2608.06663.
- Costa, F. F. (2026). Frontier AI Forecasting Has a Measurement Problem: An Audit of Progress Evidence. arXiv:2608.14903. doi:10.48550/arXiv.2608.14903.
- Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B*, 34(2), 187-202. doi:10.1111/j.2517-6161.1972.tb00899.x.
- Dellibarda Varela, I., Romero-Sorozabal, P., Rocon, E., & Cebrian, M. (2025). Rethinking the Illusion of Thinking. arXiv:2507.01231. doi:10.48550/arXiv.2507.01231.
- Du, Y., Tian, M., Ronanki, S., Rongali, S., Bodapati, S., Galstyan, A., et al. (2025). Context Length Alone Hurts LLM Performance Despite Perfect Retrieval. Findings of EMNLP 2025. arXiv:2510.05381. doi:10.48550/arXiv.2510.05381.
- Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., et al. (2023). Faith and Fate: Limits of Transformers on Compositionality. arXiv:2305.18654. doi:10.48550/arXiv.2305.18654.
- Elbers, C., & Ridder, G. (1982). True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model. *The Review of Economic Studies*, 49(3), 403-409. doi:10.2307/2297364.
- Heckman, J., & Singer, B. (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica*, 52(2), 271-320. doi:10.2307/1911491.
- Ho, A., Denain, J.-S., Atanasov, D., Albanie, S., & Shah, R. (2025). A Rosetta Stone for AI Benchmarks. arXiv:2512.00193. doi:10.48550/arXiv.2512.00193.
- Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282), 457-481. doi:10.1080/01621459.1958.10501452.
- Kapoor, S., Kirgis, P., Schwartz, A., Rabanser, S., Allaire, J. J., Bommasani, R., et al. (2026). Open-World Evaluations for Measuring Frontier AI Capabilities. arXiv:2605.20520. doi:10.48550/arXiv.2605.20520.
- Khan, S., Madhavan, S., & Natarajan, K. (2025). A Comment On "The Illusion of Thinking": Reframing the Reasoning Cliff as an Agentic Gap. arXiv:2506.18957. doi:10.48550/arXiv.2506.18957.
- Khanal, A., Tao, Y., & Zhou, J. (2026). Beyond pass@1: A Reliability Science Framework for Long-Horizon LLM Agents. arXiv:2603.29231. doi:10.48550/arXiv.2603.29231.
- Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., et al. (2025). Measuring AI Ability to Complete Long Software Tasks. *NeurIPS 2025*. arXiv:2503.14499. doi:10.48550/arXiv.2503.14499.
- Laban, P., Hayashi, H., Zhou, Y., & Neville, J. (2025). LLMs Get Lost In Multi-Turn Conversation. arXiv:2505.06120. doi:10.48550/arXiv.2505.06120.
- Lawsen, A. (2025). Comment on The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. arXiv:2506.09250. doi:10.48550/arXiv.2506.09250.
- Liu, F., Gala, J., Nilaksh, Bahdanau, D., Reddy, S., & Larochelle, H. (2026). BRIDGE: Predicting Human Task Completion Time From Model Performance. *ICML 2026*. arXiv:2602.07267. doi:10.48550/arXiv.2602.07267.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*. arXiv:2307.03172. doi:10.48550/arXiv.2307.03172.
- Mahendrakar, P. (2026). Errors That Should Compound, and When They Do: The Product Rule Assumes Independent Steps, Conformal Guarantees Assume Exchangeable Instances, and the Unit of Independence Decides What Each Can Claim. *Life of Research*. doi:10.5281/zenodo.23086063.
- Meyerson, E., Paolo, G., Dailey, R., Shahrzad, H., Francon, O., Hayes, C. F., et al. (2025). Solving a Million-Step LLM Task with Zero Errors. arXiv:2511.09030. doi:10.48550/arXiv.2511.09030.
- Mittal, S. (2026). How Fast Do Agents Rot? An Empirical Study of Long-Horizon Degradation in LLM Agents for Production Decision-Making. arXiv:2609.01660. doi:10.48550/arXiv.2609.01660.
- Ord, T. (2023). The Lindy Effect. arXiv:2308.09045. doi:10.48550/arXiv.2308.09045.
- Ord, T. (2025). Is there a half-life for the success rates of AI agents? arXiv:2505.05115. doi:10.48550/arXiv.2505.05115.
- Owen, D. (2024). How predictable is language model benchmark performance? arXiv:2401.04757. doi:10.48550/arXiv.2401.04757.
- Pai, D. M., & Xian, L. (2026). Long-Horizon State Tracking in LLMs: Executing MD5 through a Deep Sequence of Dependent Tool Calls. arXiv:2609.00012. doi:10.48550/arXiv.2609.00012.
- Pereira, D., & Zuidema, W. (2026). Transformers Struggle to Use Their Emergent World Models: Revisiting the Tower of Hanoi, and the Illusion of Thinking. arXiv:2608.07077. doi:10.48550/arXiv.2608.07077.
- Pimpale, G., Hojmark, A., Scheurer, J., & Hobbhahn, M. (2025). Forecasting Frontier Language Model Agent Capabilities. arXiv:2502.15850. doi:10.48550/arXiv.2502.15850.
- Potash, P. (2026). Playing log(N)-Questions over Wikipedia Abstracts: How Per-Round Errors Compound Under Information Asymmetry. arXiv:2609.19113. doi:10.48550/arXiv.2609.19113.
- Proschan, F. (1963). Theoretical Explanation of Observed Decreasing Failure Rate. *Technometrics*, 5(3), 375-383. doi:10.1080/00401706.1963.10490105.
- Rein, D., Becker, J., Deng, A., Nix, S., Canal, C., O'Connell, D., et al. (2025). HCAST: Human-Calibrated Autonomy Software Tasks. arXiv:2503.17354. doi:10.48550/arXiv.2503.17354.
- Ross, S., Gordon, G. J., & Bagnell, J. A. (2011). A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. *AISTATS 2011*. arXiv:1011.0686. doi:10.48550/arXiv.1011.0686.
- Ruan, Y., Maddison, C. J., & Hashimoto, T. (2024). Observational Scaling Laws and the Predictability of Language Model Performance. *NeurIPS 2024*. arXiv:2405.10938. doi:10.48550/arXiv.2405.10938.
- Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? arXiv:2304.15004. doi:10.48550/arXiv.2304.15004.
- Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *NeurIPS 2025*. arXiv:2506.06941. doi:10.48550/arXiv.2506.06941.
- Sinha, A., Arun, A., Goel, S., Staab, S., & Geiping, J. (2025). The Illusion of Diminishing Returns: Measuring Long Horizon Execution in LLMs. *ICLR 2026*. arXiv:2509.09677. doi:10.48550/arXiv.2509.09677.
- Tran-Truong, P. T., & Le, X.-B. (2026). Measuring the Unmeasurable: Markov Chain Reliability for LLM Agents. arXiv:2604.24579. doi:10.48550/arXiv.2604.24579.
- Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3), 439-454. doi:10.2307/2061224.
- Wang, X. J., Bai, H., Sun, Y., Wang, H., Zhang, S., Hu, W., et al. (2026). The Long-Horizon Task Mirage? Diagnosing Where and Why Agentic Systems Break. arXiv:2604.11978. doi:10.48550/arXiv.2604.11978.

- Weibull, W. (1951). A Statistical Distribution Function of Wide Applicability. *Journal of Applied Mechanics*, 18(3), 293-297. doi:10.1115/1.4010337.
- Whitfill, P., Snodin, B., & Becker, J. (2025). Forecasting AI Time Horizon Under Compute Slowdowns. arXiv:2511.19492. doi:10.48550/arXiv.2511.19492.
- Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley, T., et al. (2024). RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. arXiv:2411.15114. doi:10.48550/arXiv.2411.15114.
- Xing, H., Wang, P., Meng, B., Alhamoud, K., Li, X., Wang, J., et al. (2026). CurveShift: Is Agent Progress Scalar? Separating Level from Shape. arXiv:2608.00355. doi:10.48550/arXiv.2608.00355.
- Xu, F. F., Song, Y., Li, B., Tang, Y., Jain, K., Bao, M., et al. (2024). TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. arXiv:2412.14161. doi:10.48550/arXiv.2412.14161.
- Yao, S., Shinn, N., Razavi, P., & Narasimhan, K. (2024). tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. arXiv:2406.12045. doi:10.48550/arXiv.2406.12045.
- Zhang, M., Press, O., Merrill, W., Liu, A., & Smith, N. A. (2023). How Language Model Hallucinations Can Snowball. arXiv:2305.13534. doi:10.48550/arXiv.2305.13534.
- Zhu, K., Liu, Z., Li, B., Tian, M., Yang, Y., Zhang, J., et al. (2025). Where LLM Agents Fail and How They can Learn From Failures. arXiv:2509.25370. doi:10.48550/arXiv.2509.25370.

---

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.23240102