The Validation Crisis in AI Forecasting

Deflating AGI Capability Forecasts with Quantitative-Finance Overfitting Tools

Kacper Saks*

June 2, 2026

Abstract

Forecasts of when artificial general intelligence will arrive increasingly shape capital allocation, regulation, and where a generation of talent is placed — yet the confidence attached to them exceeds what the methods producing them can support. We argue this gap is structural, not careless: it is the predictable consequence of fitting a model to a measured window and projecting it forward without the validation discipline other quantitative fields learned to require. We import that discipline from quantitative finance, which developed it between 2014 and 2018 because capital was being lost to backtests that looked excellent and meant nothing. Applying three such tools — the deflated Sharpe ratio, the probability of backtest overfitting, and a walk-forward retrodiction protocol — we evaluate the major capability forecasts and introduce the Deflated Capability Forecast (DCF): a method that widens a forecast's stated interval by the amount its underlying methodology warrants, returning a distribution with explicit treatment of the tails in place of a point estimate carrying unearned precision. Across the forecasts where the method could be fully computed, deflation factors cluster between 1.3× and 2.0× — the stated intervals are systematically too narrow. We then turn the method on this work: before computing anything, we preregistered that the deflated Sharpe ratio applied to one landmark forecast would widen its interval by at least 2.3×. It produced 1.285×. We report the failure rather than revise the threshold — the failure lies not in the framework, which computed correctly, but in a prior set by intuition before the deflation it should have been derived from, which is the very error this work exists to identify. A discipline of honest validation is supposed to surface exactly this.

1 Introduction

The forecasts that shape how billions of dollars and a generation of careers are allocated toward artificial general intelligence share a methodological foundation that would not survive scrutiny in any mature quantitative discipline. This is not a claim about whether those forecasts are right; it is a claim about how they are made. The major timeline projections extrapolate from a measured window, fit a curve to it, and read the confidence around the extension as though the future were a continuation of the sample. Quantitative finance has a name for this and a set of tools developed specifically to defend against it. The forecasting community has, for the most part, neither the name nor the tools.

*Written in a personal capacity. Nothing in this paper reflects the position, knowledge, or proprietary work of any employer, past or present.

Our central argument is narrow and, we think, hard to dismiss: the current debate over AGI timelines suffers from the same methodological errors that quantitative finance diagnosed and partially solved between 2014 and 2018 — in-sample extrapolation, multiple testing without correction, the absence of walk-forward validation, and selection bias toward success. These are not exotic failures. They are the failures a hedge fund learns to detect before it is allowed to manage external capital, because each reliably produces a track record that looks excellent and means nothing. The tests that exposed overfitting in financial backtests can be applied, with care, to capability projections.

We are exact about what this paper claims and what it refuses to claim, because the distinction separates a methodological critique from a competing prophecy. It claims four things. First, that the methodology of the major published AGI forecasts — Aschenbrenner's 2024 projection [1], Cotra's 2020 and 2022 biological-anchors work [7, 8], Davidson's 2023 takeoff model [11], METR's 2023–2025 benchmark trajectories [24], and Grace's expert surveys [15, 16, 30] — is insufficient relative to the standards demanded in other forecasting disciplines. Second, that the same statistical tests which revealed overfitting in hedge-fund backtests [2, 3] can be sensibly applied to these projections. Third, that production deployment of advanced AI in safety-critical systems requires a validation infrastructure current timelines underweight or omit [14, 20]. Fourth, that the European regulatory stack [13] introduces structural friction absent from US-centric forecasts that materially affects any deployment timeline.

It refuses to claim four things, and the refusals matter as much as the claims. This paper does not argue that AGI will not arrive. It does not argue that the forecasters are dishonest or incompetent — it critiques methodology, not people, and cites the work without a single pejorative adjective, because the work is serious and the seriousness is exactly why its methodological exposure is worth examining. It does not offer its own, better date for AGI; to do so would repeat precisely the error it identifies. And it does not claim that quantitative finance has solved its own validation problems — only that finance has paid, in real capital and over real years, for a discipline that capability forecasting has not yet adopted.

The structure recurs across the literature: a window, a fit, an extension, and a confidence interval that has not been deflated for the number of model variations implicitly searched in producing it. Aschenbrenner decomposes recent progress into orders of magnitude of effective compute and extends the rate forward; Cotra estimates the training compute transformative AI would require against six biological anchors spanning seventeen orders of magnitude, and the 2022 update's twelve-year revision of its own median in two years is the most informative event in the framework's history; Davidson exposes roughly seventy parameters whose one-at-a-time sensitivity analysis is, under multiple-testing logic, an exposure rather than a defense. METR and Grace are the strongest of the set on the dimension where the others are weakest — both rest on genuinely measured quantities — and they fail the same test anyway: METR's doubling rate is real data, but the projection two-to-five years beyond it extrapolates a window-dependent slope; Grace's surveys are the densest record of expert opinion in the field, yet produce median timelines that swing by sixty to a hundred and five years on the framing of the question. Explicitness about an in-sample boundary is not the same as a protocol for crossing it.

The tool that does the most work is walk-forward validation, which manufactures the

held-out sample a forecast never reserved: it reconstructs the forecast as of its publication date, freezes the information set to what was then available, and scores its projection against the period that has since elapsed. Several surveyed forecasts have been public long enough that this held-out sample exists; it simply has not been used. We are careful that the adaptations are not free — the financial deflation assumes a stationary, independent series, capability progress is neither; the number of configurations searched is rarely disclosed; the performance statistic must be constructed for the domain rather than borrowed — and Section 2 treats each as a derivation problem, marking provisional any result whose required assumption cannot yet be validated.

Section 2 develops this method and constructs the Deflated Capability Forecast (DCF): it takes a forecast's point estimate and too-narrow interval and widens the interval by the amount the underlying methodology warrants, returning a distribution with explicit treatment of the tails in place of unearned precision. Section 3 applies it. Here the accounting must be honest about its own scope: the framework is computed in full for those forecasts where the necessary data could be reconstructed — Aschenbrenner, both Cotra configurations, and Davidson — while METR and Grace are surveyed here but not fully deflated, their per-anchor and per-respondent records remaining open dependencies that Section 3 marks as provisional rather than reporting a factor it has not earned. To these computed forecasts Section 3 adds one more: the author's own preregistered self-prediction, entered as the integrity capstone. Section 4 turns from statistics to deployment — the production-reality gap and the European regulatory stack — and to the interpretation of that self-application.

2 Method

The framework is the quantitative-finance overfitting canon adapted to capability projection: in-sample extrapolation, walk-forward retrodiction, multiple-testing correction, the Deflated Sharpe Ratio (DSR), the Probability of Forecast Overfitting (PFO), the effective-trial-count composite both adaptations require, and the Deflated Capability Forecast (DCF) that composes them into a distribution over capability-deployment time. Three honesty statements close the section. Full derivations are in chapters 4 through 7 and 14 of the extended treatment [28]; this section states the equations and structural moves.

2.1 In-sample extrapolation

In-sample extrapolation measures a quantity over a historical window, fits a model to that window, and projects forward as though in-sample conditions hold out-of-sample. The confidence interval around an in-sample fit is almost always too narrow because it does not account for the search that produced the fit. The mechanism is selection bias under multiple testing: when many configurations are evaluated and the in-sample best is reported, the reported statistic is the maximum over configurations searched, and the sampling distribution of a maximum sits above that of a single trial. The narrowness scales with the number of configurations searched before the best is reported [2, 4].

The capability forecasts examined here repeat the move. Each reconstructs a measured window — orders of magnitude of effective compute, biological-anchor mixtures, takeoff-duration parameter couplings, benchmark-task time horizons, expert-survey waves — fits

a model, and extends forward without a held-out criterion. A forecast that reports its in-sample fit as evidence about an unobserved future window faces the same in-sample-fitting risk a backtest does, and the same correction-for-search- multiplicity applies [28].

2.2 Walk-forward retrodiction

Walk-forward retrodiction manufactures a held-out sample the forecast did not reserve. The forecast is reconstructed as of its publication date with the information set frozen to what was then available; its projection is scored against the period that has since elapsed. The in-sample window is frozen at the publication vintage; the elapsed post-publication period is the test region; the boundary is purged of leakage so that no post-publication observation contaminates the reconstructed fit [23, ch. 7]. The protocol applies only where a test region has elapsed; the five forecasts named in Section 1 each carry elapsed projection windows of at least one year.

2.3 Multiple-testing correction

The multiple-testing correction responds to the maximum-over-N inflation of Section 2.1. The graded lineage runs from family-wise error control [19] through false-discovery-rate control [5] to the finance-domain application of Harvey, Liu, and Zhu [17]: catalogued across hundreds of published predictors of cross-sectional returns, the corrected significance threshold rises substantially above the conventional one, and a large share of published predictors do not survive it. The diagnosis transfers to the capability domain as a question — how many decomposition specifications the forecaster searched, and how many of those the published methodology discloses — with the decomposition-search count as the trial count the correction acts on and the disclosed-versus-unenumerated status determining the adjustment's magnitude [28].

2.4 Deflated Sharpe Ratio (Adaptations 1, 3)

The Deflated Sharpe Ratio (DSR) is derived in three steps [2]. The first is the sampling distribution of the performance statistic: under independent and identically distributed returns with finite second moments, the estimated Sharpe ratio is asymptotically normal, with variance σ^2_{SR} a function of the true Sharpe, the skewness γ_3, and the excess kurtosis γ_4 [22]. The second is the distribution of the maximum: when a researcher searches N candidate configurations and reports the best in-sample Sharpe, the reported statistic is the maximum of N draws, and under the null $\mathrm{SR} = 0$ the expected maximum is the expected value of the largest of N standard normals, with closed-form approximation $\mathbb{E}[\mathrm{SR}_{\max}]$ reducing for large N to $\sqrt{2\ln N}/\sqrt{T}$ [2, eq. 5]. The third is the deflation itself:

$$\mathrm{DSR} = \Phi\left(\frac{(\mathrm{SR}_{\max} - \mathbb{E}[\mathrm{SR}_{\max}])\sqrt{T-1}}{\sqrt{\sigma^2_{\mathrm{SR}}}}\right), \tag{1}$$

with the Lo [2002] standard error in the denominator, finite-sample adjusted, written in the excess-kurtosis convention used throughout ($\gamma_4 = 0$ for a Gaussian).

Three formula components do not transfer unaltered. Adaptation 1 replaces the IID variance term with a stationary-bootstrap estimator [26], because capability-progress series are non-stationary and serially dependent. Adaptation 3 constructs the performance statistic the variance is a property of: the financial Sharpe ratio has no off-the-shelf capability analog, so a five-candidate panel — Brier score, log-loss, calibration error, interval coverage, capability-Sharpe on doubling-rate residuals — is selected per forecast type with the selection preregistered [28]. Adaptation 4 (effective-trial- count composite) is shared with the PFO and treated in Section 2.6.

2.5 Probability of Forecast Overfitting (Adaptation 2)

The Probability of (Backtest) Overfitting (PBO) of Bailey et al. [3] asks a different question from the DSR. The DSR corrects the reported statistic downward for the expected best-of-N; the PBO estimates, across the configurations actually searched, the frequency with which the in-sample-best configuration falls below the out-of-sample median. The record is partitioned into S equal blocks; the $\binom{S}{S/2}$ ways of assigning $S/2$ blocks to the in-sample half and $S/2$ to the out-of-sample half are enumerated; for each split, the configuration with the maximum in-sample rank is identified and its out-of-sample logit $\lambda_s = \log(w_s/(N - w_s + 1))$ recorded; the overfitting probability is the frequency of splits in which the in-sample-best lands below the out-of-sample median:

$$\text{PBO} = \frac{1}{\binom{S}{S/2}} \sum_s \mathbf{1}\{\lambda_s < 0\}. \tag{2}$$

PBO ranges in $[0, 1]$; PBO $= 0.5$ is the no-information baseline; PBO > 0.5 the regime in which in-sample selection is actively misleading.

Combinatorial symmetry rests on exchangeable strategy rankings across the partition. Capability forecasts published at different dates are not exchangeable: a later forecast has access to evidence an earlier one did not, so the symmetric partition cannot be imported directly. Adaptation 2 replaces the combinatorially-symmetric partition with an information-set-respecting sequential-test partition under the publication-time filtration. For consecutive publication times, the in-sample set is the data observable through the earlier date, the out-of-sample set is the increment between them, and the per-test statistic asks whether the forecast that ranked first under the earlier information set underperforms in the increment its publication date held out. The aggregate Probability of Forecast Overfitting (PFO) is the frequency of sequential tests in which it does; the no-forecasting-skill null is PFO $= 0.5$, the same logit-transform reporting and bootstrap confidence intervals as the canonical PBO [28]. The reformulation has two failure modes the symmetric partition produces in the capability domain: blocks observable only after a forecast's publication assigned to that forecast's in-sample half, and forecasts published after the median date inheriting capability data available at the median — either failure returns a number that does not measure forecast overfitting.

2.6 Effective-N composite (Adaptation 4)

The effective-trial-count N drives both $\mathbb{E}[\mathrm{SR}_{\max}]$ in the DSR and the multiplicity correction in the PFO. The capability domain departs from the finance regime along three structural axes: the search space is bounded but only partially disclosed (Axis A); the search is distributed across a research community with no central log, the file-drawer problem of Rosenthal [27] (Axis B); and later forecasts inherit modeling choices from earlier ones, so cumulative community N exceeds individual-forecast N (Axis C). No single estimator addresses all three. Adaptation 4 constructs four candidate algorithms: Algorithm D, the sensitivity-disclosure floor (a strict lower bound from the forecaster's published variant count); and Algorithms A, B, C — a citation-network proxy, a publication-rate proxy, and a compute-budget proxy — upward-biased estimates of community search depth. The composite range estimator is

$$N_{\text{range}} = \big[\, \max(N_D),\ \min(\max(N_A, N_B, N_C)) \,\big], \tag{3}$$

with the headline reported as the geometric mean $\sqrt{N_{\text{lower}} \cdot N_{\text{upper}}}$, preferred over the arithmetic because the deflation's sensitivity to N is logarithmic — $\mathbb{E}[\mathrm{SR}_{\max}]$ scaling with $\sqrt{2 \ln N}$ — so geometric averaging on the count scale corresponds to arithmetic averaging on the $\log N$ scale where the deflation varies linearly. Two preregistered thresholds govern reporting: where the upper-to-lower bracket ratio exceeds $100\times$ or falls below $1.5\times$, the result is reported as a range without headline. A pre-committed bias-direction anomaly rule covers the pathological case: if the disclosure floor exceeds the minimum upward-biased proxy, the result is reported as anomalous and downgraded [28].

Per-axis N_{upper} values consumed by Section 3 are computed in chapter 7 of the extended treatment from disclosure-floor counts and the four-algorithm composite.

2.7 The Deflated Capability Forecast — master equation

The Deflated Capability Forecast composes three corrections — the DSR deflation of Section 2.4, the PFO weighting of Section 2.5, and the certification-friction factor synthesized from the regulatory-stack analysis [28] — into a distribution over capability-deployment time. The order is fixed: deflate the reported point estimate for the implicit search behind it, weight the deflated estimate by the overfitting probability, then widen and shift the resulting interval by the certification-friction factor (φ, δ), where $\varphi \geq 1$ is the multiplicative widening converting capability-arrival half-width to deployment half-width and $\delta \geq 0$ is the activity-weighted regulatory lead time. The first two corrections operate on the capability-arrival interval; the third maps that interval onto a deployment interval.

Writing the reported interval as $[t_{\text{lo}}, t_{\text{hi}}]$ with center $\hat{t}$ and half-width $w = (t_{\text{hi}} - t_{\text{lo}})/2$, the deflated center is $\hat{t}_D = \hat{t} + a \cdot d \cdot w$ (with a defaulting to 1 and $d = 1 - \mathrm{DSR}$ the dimensionless deflation magnitude), and the deflated half-width is $w_D = w \cdot (1 + b \cdot d)$ (with b defaulting to 1). Writing Φ^{-1} for the inverse standard-normal CDF, $z_\alpha = \Phi^{-1}(1-\alpha)$, and $H = w_D \cdot (1 + \kappa \cdot \max(0, \mathrm{PFO} - \frac{1}{2}))$ the rectified deflated half-width, the DCF distribution is

Equation 14.1' (master).

$$F_{\mathrm{DCF}}(t) = \begin{cases} F_{\mathrm{spine}}(t) & t \le u_R, \\ (1-\alpha) + \alpha \cdot F_{\mathrm{GPD}}(t - u_R \mid \sigma_R, \xi) & t > u_R, \end{cases} \tag{4}$$

with the structural left-tail report convention $\Pr(T < \hat{t}_D - w_D) \equiv 0$,

where the **spine** is $F_{\mathrm{spine}} = \mathrm{Normal}(\mu, \sigma^2)$ with matching condition

$$\mu = \hat{t}_D + \delta, \qquad \sigma = \varphi \cdot H / z_\alpha, \qquad z_\alpha = \Phi^{-1}(1-\alpha),$$

so that the spine's $(1-2\alpha)$-central interval reduces under the matching condition to $[\mu - \varphi H,\ \mu + \varphi H]$, the symmetric corollary stated below; the **right-tail layer** is the generalized-Pareto CDF of Pickands [25] with scale σ_R and shape ξ, spliced at $u_R = \mu + \varphi H$ so that $F_{\mathrm{DCF}}(u_R) = 1 - \alpha$ (continuity of mass at the splice); and the **left-tail floor** $\Pr(T < \hat{t}_D - w_D) \equiv 0$ is the deflation-discipline structural report (a capability cannot be deployed before the date on which its skill is established). The right-tail parameters are parameterized from the activity-weighted regulatory-friction mechanism: $\sigma_R = \sum_r a_r \cdot \delta_r$, $\xi = \max_r \xi_r$ with default $\xi = 0$ (the Gumbel/exponential limit Cramér 10 pins under regularity).

The symmetric corollary Section 3 uses as the operative reporting object falls out of Equation 4 as the spine's $(1-2\alpha)$-central interval:

Equation 14.1 (corollary).

$$\mathrm{DCF}_{\mathrm{CI}}([t_{\mathrm{lo}}, t_{\mathrm{hi}}]) = \Big[\hat{t}_D + \delta - \varphi \cdot w_D \cdot (1 + \kappa \cdot \max(0,\, \mathrm{PFO} - \tfrac{1}{2})), \\ \hat{t}_D + \delta + \varphi \cdot w_D \cdot (1 + \kappa \cdot \max(0,\, \mathrm{PFO} - \tfrac{1}{2})) \Big]. \tag{5}$$

The κ-rectification — replacing a signed $(\mathrm{PFO} - \frac{1}{2})$ with the rectified $\max(0, \mathrm{PFO} - \frac{1}{2})$, with κ held at the parsimonious constant $\kappa = 1$ — is the derived correction. Its role is to ensure the PFO term can only widen the spine, never narrow it: a PFO below the no-information baseline of $\frac{1}{2}$ would otherwise let overfitting-statistic noise spuriously certify a forecast as more precise than it reported, an asymmetry with no warrant. At $\mathrm{PFO} = \frac{1}{2}$ the rectified term is exactly zero and the κ coefficient is inert; above $\frac{1}{2}$ it activates and widens [28].

The components: $\hat{t}_D$ is the deflated center and w_D the deflated half-width (Section 2.4); PFO is the Probability of Forecast Overfitting at the headline effective N (Section 2.5); $\kappa \ge 0$ is the overfitting-weighting coefficient, fixed at $\kappa = 1$; $\varphi \ge 1$ and $\delta \ge 0$ are the certification-friction factor (multiplicative widening and activity-weighted regulatory lead time of the regime decomposition); σ_R and ξ are the EVT right-tail parameters of the residual friction-distribution shape above u_R [10, 25]. The no-double-count property holds because the channels are orthogonal: d enters only through w_D and the center shift; PFO only through the spine half-width; φ scales H; δ shifts the location; (σ_R, ξ) parameterize the tail shape above the mean. The full derivation and the dimensional analysis are in chapter 14 of the extended treatment [28].

2.8 Asymmetric tail treatment

The asymmetric right-tail treatment is required by the directional asymmetry of the regulatory-friction mechanism: regulatory delay only lengthens the deployment date, never shortens it, so a symmetric Gaussian alone misrepresents the deployment-date distribution. Cramér's [1946] extreme-value limit theorem pins the Gumbel/exponential regime under regularity ($\xi = 0$); Pickands [25] supplies the generalized-Pareto distribution as the peaks-over-threshold limit, parameterizing the right tail above u_R with scale σ_R and shape ξ. The EVT layer prices the directional friction the symmetric interval cannot. Full algebraic derivation — the GPD-splice continuity proof, the no-double-count argument for the spine's φ versus the tail's σ_R, the verification that the master reduces to the corollary at R1 (symmetric-Gaussian limit), to the canonical DSR/PBO at R2 ($\varphi = 1$, $\delta = 0$, PFO $= \frac{1}{2}$), and to the reported interval at R3 (additionally $d = 0$, the genuine-skill limit) — is offloaded to chapter 14 §4 of the extended treatment and exercised by the reference notebooks [28].

2.9 The adaptations are not free

The transfer of the finance canon to the capability domain rests on three concessions, each carried as an explicit honesty statement and each shaping the epistemic labeling of the magnitudes Section 3 reports. First, **non-stationarity**: the canonical DSR variance σ^2_{SR} rests on the IID assumption, and capability-progress series violate it for three structural reasons — secular trends in training compute [29], algorithmic-efficiency improvement [18], and autocorrelated capability clusters. Adaptation 1 replaces the Lo [2002] IID variance with the stationary-bootstrap estimator of Politis and Romano [26], but the consistency result is empirically validated only at sample sizes at or above 64, and the surveyed capability series sit well below that floor; Adaptation 1 is therefore labeled *(speculative)* by default.

Second, **undisclosed search count**: the number of configurations searched, which drives the deflation's magnitude, is rarely disclosed and in several cases not enumerable from the published methodology. Adaptation 4 (Section 2.6) ranges the effective N rather than asserting a point, and where the bracket ratio falls outside the preregistered $[1.5\times, 100\times]$ bounds the result is reported as a range without headline. The composite is *(speculative)* by default; its elevation requires empirical validation studies absent from the current literature.

Third, **missing performance statistic**: the financial Sharpe ratio has no unique capability-forecasting analog, so Adaptation 3 evaluates a five-candidate panel (Brier score, log-loss, calibration error, interval-coverage statistic, capability-Sharpe analog on doubling-rate residuals) and selects per forecast type, with the per-framework selection preregistered. Asymptotic distributions for the canonical scoring rules are primary-source-attested *(evidenced)*; the capability-Sharpe analog remains *(speculative)* in the small-sample and heavy-tailed regimes common to the surveyed series.

The DCF inherits every one of those labels. The canonical components — Lo [22], Bailey and López de Prado [2], Bailey et al. [4], Cramér [10], Pickands [25] — are *(established)*; the capability adaptations carry their *(speculative)* defaults; the composition itself is *(speculative — derived)* because the composed object has not been validated against a realized capability outcome.

2.10 Provisional cells discipline

Several per-forecast cells in Section 3 are reported as provisional because their input series are not yet recovered. The provisional labels distinguish a magnitude the framework returns from a magnitude pending data the framework requires. Each provisional cell carries a *(speculative — pending D-N ...)* marker naming the dependency — per-anchor Brier extraction, doubling-rate residual construction, per-respondent survey data, OOM residual-series construction, and the compute-scaling residual protocol. Section 3 reports what the framework computes at this revision's data state.

3 Results

The framework of Section 2 is applied to the five surveyed forecasts of Section 1 — Aschenbrenner (2024), Cotra (2020), Cotra (2022), Davidson (2023), and the document's own preregistered HYPOTHESIS_A_CONFIRMED — and the headline result is reported in a single five-row deflation table. The per-forecast subsections of 3.1 through 3.5 fix the inputs the `deflated-capability-forecast` package consumes and state the magnitudes the package returns; the consolidated table of Section 3.6 reports the headline width ratios alongside Figure 1, which renders the original 95% reported intervals against the deflated DCF intervals row by row; the closing reading of Section 3.7 records that no forecast in the surveyed set reached the preregistered $\geq 2.3\times$ threshold. Every numerical magnitude reproduced here matches the values pinned in chapter 16 of the extended treatment [28] and is regenerated to a 10^{-3} tolerance in the accompanying reproducibility repository.

3.1 Aschenbrenner (OOM-axis, 2027)

The Aschenbrenner inputs consumed by Equation 4 are the Part-1 reconstruction at chapter 1.1 of the extended treatment: a 2027 central reading from the four-OOM additive decomposition; an operationalized reported interval $[2025.000, 2029.000]$ (the symmetric four-year band the per-OOM driver-sum range maps onto at approximately 1.25 OOM/year, centered on the 2027 headline); the chapter 7 §3 effective-N composite $N_{\text{lower}} = 7$, $N_{\text{upper}} = 189$, $N_{\text{headline}} = 36.39$ (C-only scope, bracket PASS at $27.0\times$); methodology characterization as parametric extrapolation with interval-coverage Adaptation 3 (Resolution C) and the sequential-test partition of Adaptation 2; and the certification-friction aggregate $(\varphi, \delta, \sigma_R, \xi) = (1.1975, 0.3400, 0.3400, 0)$ of chapter 13 §4. Stage 1 returns DSR $= 0.7151$ and $d = 0.2849$; Stage 2's overfitting factor at the null-reference PFO $= 0.500$ is 1.000 and the spine half-width is unwidened by the rectified term; Stage 3 shifts the center by $\delta = 0.34$ years and scales the spine by φ.

The full Equation 4 application returns a DCF central interval $[2024.833, 2030.987]$, tail-quantile triple ($P_{10} = 2025.898$, $P_{50} = 2027.910$, $P_{90} = 2029.922$), right-tail mass at $u_R = 2030.987$ of 0.025, structural left-floor mass at $t_{\text{floor}} = 2025.000$ of exactly 0, and width ratio $\mathbf{1.539}\times$. The Stage-1 DSR-only reading is distinct: the deflated CI under Equation 5 at $\varphi = 1$, $\delta = 0$, PFO $= \frac{1}{2}$ is $[2025.000, 2030.139]$, width 5.139 years against the reported 4.000 years, ratio $\mathbf{1.285}\times$.

The document preregistered, at master plan A.16 line 442 and preregistration v2, the pre-

diction that the DSR-only deflation of Aschenbrenner's OOM extrapolation would produce a 95% confidence interval at least 2.3× wider than the reported 95% CI. The framework produced 1.285×. The prediction **FAILED**. The threshold the preregistered claim names is $2025.000 + 2.3 \times 4.000 = 2034.200$; the deflated upper bound is 2030.139; the gap between the Stage-1 bar and the preregistered threshold line is **4.061 years**, the visibly unambiguous distance Figure 1 renders against the marked threshold line. The root cause is the author's calibration, not the framework's behavior: the $N_{\text{lower}} = 7$, $N_{\text{upper}} = 189$, $N_{\text{headline}} = 36$ composite places Aschenbrenner in a modest-deflation regime ($d = 0.285$), and the operationalized four-year reported band leaves the proportional widening less room to inflate than the preregistered $\geq 2.3\times$ prior assumed. The 1.539× full DCF width ratio answers a different question — the asymmetric Equation 4 with friction shift, friction widening, and the EVT right-tail layer — and does not soften the 1.285× result the Stage-1 DSR-only prediction named.

3.2 Cotra 2020 (anchor-axis, 2052)

The Cotra 2020 inputs are the chapter 1.2 §1 Part-1 reconstruction: a 2052 median TAI-arrival reading from the six-anchor mixture under the 2020 weights; reported interval $[2031.000, 2100.000]$ operationalized as the source's P10–P80 percentile pair read as a 95% reading; the chapter 7 §3 anchor-axis-2020 composite $N_{\text{lower}} = 18$, $N_{\text{upper}} = 69$, $N_{\text{headline}} = 35.32$ (Lang-scope primary, bracket PASS at 3.85×); methodology as probabilistic distribution with Brier-on-per-anchor-events Adaptation 3 (Resolution A) and sequential-test Adaptation 2; the same chapter 13 §4 certification-friction aggregate as Aschenbrenner. Stage 1 returns DSR = 0.7218 and $d = 0.2782$; Stage 2 returns the rectified overfitting factor 1.000 at PFO = 0.4857 (the rectification holds the spine half-width unchanged at and below the no-information baseline of $\frac{1}{2}$).

Equation 4 returns DCF central $[2022.630, 2128.247]$, tail-quantile triple ($P_{10} = 2040.909$, $P_{50} = 2075.439$, $P_{90} = 2109.968$), right-tail mass at u_R of 0.025, and width ratio **1.531**×. The deflated CI under Stage 1 alone is $[2031.000, 2119.198]$ — the proportional widening the larger reported band carries against its modest-deflation composite. The Cotra 2020 P_{50} of 2075.4 calendar years sits approximately 23 years past the source's reported 2052 median, the friction-shift δ plus the spine-center translation by $a \cdot d \cdot w = 0.2782 \times 34.5 = 9.6$ years jointly accounting for the displacement.

3.3 Cotra 2022 (anchor-axis alt-scope, 2040)

The Cotra 2022 alt-scope inputs differ from the 2020 reading in two structural respects. The headline median shifts from approximately 2052 to approximately 2040, the published percentile structure tightening to $P_{15} = 2030$, $P_{60} = 2050$, $P_{97} = 2100$ under the re-elicited weights and the partial TAI redefinition. The chapter 7 §3 anchor-axis-2022 composite is $N_{\text{lower}} = 18$, $N_{\text{upper}} = 296$, $N_{\text{headline}} = 72.99$ (Lang-scope primary, bracket PASS at 16.4×) — structurally larger than the 2020 reading's composite because the 2022 re-elicitation adds a second weight search to the first, increasing the disclosed decomposition-choice count chapter 5 §3 enumerates. The methodology characterization, certification-friction aggregate, and per-axis statistic selection are inherited from the 2020 reading.

Stage 1 returns DSR = 0.5534 and $d = 0.4466$ — the larger N_{upper} of the 2022 alt-scope

drives a stronger deflation than the 2020 reading at the same per-axis Brier statistic, the framework's structural correction for the second weight search the re-elicitation added. Equation 4 returns DCF central [2020.340, 2141.602], tail-quantile triple ($P_{10} = 2041.327,\ P_{50} = 2080.971,\ P_{90} = 2120.615$), right-tail mass at u_R of 0.025, and width ratio $\mathbf{1.732}\times$. The two Cotra readings are reported as two distinct deflations of two distinct published intervals at two distinct effective-N composites, not as a verdict on whether the 2022 update improved the 2020 forecast in the framework's terms.

3.4 Davidson (takeoff-duration, post-wake-up)

The Davidson inputs are the chapter 1.3 §1 Part-1 reconstruction: a three-point probability distribution over post-wake-up takeoff duration (∼25% under one year, 50% under three years, 80% under ten years); reported interval [1.000, 10.000] years operationalized as the P25–P80 percentile pair read as a 95% reading; the chapter 7 §3 takeoff-duration composite $N_{\text{lower}} = 70$, $N_{\text{upper}} = 706$, $N_{\text{headline}} = 222.24$ (Lang-only primary, bracket PASS at 10.1×); methodology as probabilistic distribution over a duration with Brier Adaptation 3 (Resolution A) and sequential-test Adaptation 2; the same chapter 13 §4 certification-friction aggregate. The takeoff-duration axis is in the seven-axis PFO scope but excluded from the chapter 6 §4 deflation-magnitude five-axis scope (no elapsed sample exists to score the takeoff distribution against — the headline event has not occurred); the DCF nonetheless applies, returning a distribution over the duration whose Stage 1 deflation acts on the at-publication uncertainty rather than on an elapsed-sample residual statistic.

Stage 1 returns DSR = 0.3122 and $d = 0.6878$ — the strongest deflation of the four external forecasts, consistent with Davidson's largest implicit hyperparameter-search space among the surveyed set: the seventy-named-parameter simulator burns through more of the multiple-testing budget than the smaller Cotra and Aschenbrenner axes do. Stage 2's overfitting factor is 1.000 at the null-reference PFO = 0.5000. Equation 4 returns DCF central [−0.160, 18.030], tail-quantile triple ($P_{10} = 2.988,\ P_{50} = 8.935,\ P_{90} = 14.882$), right-tail mass at u_R of 0.025, and width ratio $\mathbf{2.021}\times$ — the largest of the surveyed set. The nominal $P_{2.5} = -0.160$ years is the spine Gaussian's matched-scale lower quantile (the spine is $N(\mu = 8.935,\ \sigma = 4.640)$ at the matching condition $\sigma = \varphi \cdot H/z_\alpha$); the asymmetric DCF's structural left-floor convention applies at $t_{\text{floor}} = 1.000$ year (the deflated lower bound), so the realized left-tail mass below 1 year is exactly 0 (a duration cannot be negative).

3.5 HYPOTHESIS_A_CONFIRMED (self-application)

The document's own preregistered HYPOTHESIS_A_CONFIRMED — locked at SHA-256 `a2f7e52c...` per preregistration v1 §"Distribution analysis" — is cast into Equation 4 input form per the Editor-in-Chief's sub-decision 1 (continuous-quantile reading). The headline narrative-philosophical 38% is the P_{50} location; the forecast-philosophy bracket [10–27%] is the left-mass anchor; the empirical-measurement bracket [40–82.5%] is the right-mass anchor. The cast operationalizes the reported interval as [10.000, 82.500] on the *(established)*-share %-scale (the widest defensible 95% reading of the EiC's sub-decision-1 anchors). The composite is $N_{\text{lower}} = 7$ (the seven-instance training set), $N_{\text{upper}} = 21$ (seven instances × three category labels), $N_{\text{headline}} = 12.12$ (bracket PASS at 3.00×). The methodology characterization is probabilistic distribution over the *(established)*-share quantity with Brier

Adaptation 3 (Resolution A) and sequential-test Adaptation 2. The certification-friction factor is null — HYPOTHESIS_A is a meta-methodological prediction, not a deployment-domain quantity — so $(\varphi, \delta, \sigma_R)$ collapse to $(1, 0, 0)$ and the EVT layer reduces to the spine Gaussian's nominal upper tail with $\xi = 0$. Stage 1 returns DSR = 0.6797 and $d = 0.3203$; Stage 2's overfitting factor is 1.000 at the null-reference PFO = 0.5000.

Equation 4 returns DCF central $[10.000, 105.719]$ on the %-scale, tail-quantile triple ($P_{10} = 26.566$, $P_{50} = 57.860$, $P_{90} = 89.153$), right-tail mass at $u_R = 105.719$ of 0.025, structural left-floor mass at $t_{\text{floor}} = 10.000$ of 0, and width ratio $\mathbf{1.320}\times$. The deflated CI equals the DCF central CI exactly because the friction layer is null: the asymmetric Equation 4 collapses to the spine Gaussian's own upper tail with no EVT extension above u_R. The asymmetric DCF applied to a $[0, 100\%]$-bounded variable produces a right tail ($P_{97.5} = \mathbf{105.719}\%$) exceeding the bound; the cast reports the raw value and flags the bound rather than clipping, since clipping would understate the deflation the method computes. The conservative-honesty cost the continuous-domain cast inherits is the spine Gaussian's nominal upper tail above 100%, reported at the raw cast value and reproduced in Figure 1 past the $[0, 100\%]$ dotted boundary.

3.6 Consolidated deflation table

The five-row table (Table 1) consolidates the per-forecast magnitudes of Sections 3.1 through 3.5 at the precision pinned in chapter 16 §6 of the extended treatment. Each row is one application of the framework Section 2 constructed to one published forecast input set, computed by the `deflated-capability-forecast` package and re-verified to a 10^{-3} tolerance in the accompanying reproducibility repository.

Table 1: Five-forecast deflation table. DCF Stage 1 is the DSR-only corollary (Equation 5 at $\varphi = 1$, $\delta = 0$, PFO = $\frac{1}{2}$); Stage 2 is the full master Equation 4. Stage-1 cells for four non-Aschenbrenner rows marked *(speculative — pending D-N…)* are provisional because their input series are not yet recovered (D-2: per-anchor Brier extraction; D-3: doubling-rate residual construction).

Forecast	**Reported 95% CI**	**Stage 1 ratio**	**Stage 2 ratio**	**Notes**
Aschenbrenner (OOM, 2027)	$[2025.0, 2029.0]$ yr	$\mathbf{1.285}\times$	$\mathbf{1.539}\times$	**FAILED** vs. preregistered $\geq 2.3\times$; gap = 4.061 yr to threshold
Cotra 2020 (anchor, 2052)	$[2031.0, 2100.0]$ yr	*(spec. pend. D-2)*	$\mathbf{1.531}\times$	$P_{50} = 2075.4$ yr
Cotra 2022 (anchor alt., 2040)	$[2030.0, 2100.0]$ yr	*(spec. pend. D-2)*	$\mathbf{1.732}\times$	$P_{50} = 2081.0$ yr; 2nd-iter. weight search drives stronger deflation
Davidson (takeoff, post-wake-up)	$[1.0, 10.0]$ yr	*(spec. pend. D-3)*	$\mathbf{2.021}\times$	$P_{50} = 8.9$ yr; left-floor at $t_{\text{floor}} = 1.0$ yr
HYP-A CONFIRMED (self)	$[10.0, 82.5]\%$	—	$\mathbf{1.320}\times$	$P_{97.5} = \mathbf{105.719}\%$ unclipped; right tail exceeds $[0, 100\%]$ bound

Figure 1 is reproduced in the accompanying reproducibility repository; every plotted value is re-verified against the `deflated-capability-forecast` package to a 10^{-3} tolerance, and the rendered visual properties (the Aschenbrenner DSR-only bar falling short of the marked 2.3× threshold line; the HYPOTHESIS_A bar extending past 100% unclipped) are asserted programmatically. Full provenance, the per-row pinned-value trace, and the asymmetric-derivation reference are included therein.

3.7 The 1.28×–2.02× pattern

The width ratios of Section 3.6 cluster between **1.285**× (Aschenbrenner DSR-only Stage 1) and **2.021**× (Davidson full Equation 4) across the surveyed five computes (Aschenbrenner DSR-only Stage 1 at 1.285×; Aschenbrenner full Stage 2 at 1.539×; Cotra 2020 at 1.531×; Cotra 2022 alt-scope at 1.732×; Davidson at 2.021×; HYPOTHESIS_A_CONFIRMED self-application at 1.320×). No forecast hit the $\geq 2.3\times$ threshold the document's preregistered prior named at master plan A.16. The framework produced a uniform 1.28×–2.02× deflation range; the preregistered self-prediction on Aschenbrenner ($\geq 2.3\times$ DSR-only Stage 1) FAILED at 1.285×.

The pattern is the chapter's strongest finding. It is not an Aschenbrenner-row quirk; it is a calibration failure of the author's preregistered prior about what the framework would produce, surfaced by the framework's operation across the five-forecast surveyed set, and reported here under the same epistemic labels and at the same precision as the four external-forecast applications it accompanies. The framework deflates real but less dramatically than the author's prior assumed. Section 4 picks up the broader reading; Section 3 establishes the empirical facts.

4 Discussion

The deflation magnitudes of Section 3 are statistical objects; the distance between a capability statistic and a deployed capability is institutional. This section reads the surveyed deflations against the deployment-side context the framework's certification-friction inputs come from — the production-reality gap [28] and the European regulatory stack [13] — and against the broader self-application pattern. The full treatments are in the chapters cited; this section names the structural arguments and the limitations the deflations inherit.

4.1 Production reality gap

A capability claim that clears the Section 2 statistical bar has demonstrated only that the capability plausibly exists out-of-sample; it has not demonstrated that an institution with legal authority over a deployment domain can validate that the capability performs to a stated standard under the conditions of use, and the second is the work of validation infrastructure. The two are distinct in kind, not degree, and the major published timelines forecast the first while being read as forecasting the second. That conflation is the structural reason a capability-arrival date and a deployment date are different quantities [28].

The validation discipline Section 2 made statistical is the same discipline certification regimes have institutionalized for decades. Three families of regime ground the argument.

Industrial-automation functional safety indexes its evidence burden through Safety Integrity Levels and Performance Levels fixed by the consequence of failure, and any modification to a safety-related system re-opens the evidence body — a more capable AI component embedded in a safety function is a modification, not a discharge [20, 28]. Regulated finance institutionalizes the same demand through model-risk governance and an independent validation cycle that runs on the supervisor's calendar, not the model's training calendar [6, 28]. Medical-device AI carries the sharpest publicly codified example through the FDA's Predetermined Change Control Plan, which requires a manufacturer to pre-specify, at authorization, the anticipated capability changes and the methodology by which each will be validated — the same ex-ante discipline the DCF and this document's own preregistration embody [14, 28]. The capability–certification gap is the formal counterpart of these instruments: the certification standards are not a separate objection to the timeline forecasts but the same discipline, already institutionalized where an unvalidated claim is paid for in lives and capital.

The bottleneck does not dissolve as the model improves, because the certification-evidence burden was never a function of the model's score. The institutional capacity to certify is bounded by inputs independent of capability — qualified auditors, regulatory-body throughput, accumulated operational data — and a faster capability doubling widens the gap rather than narrowing it [28]. The two shared properties making each regime indifferent to capability improvement are consequence-indexed evidence and structural independence between the party that confirms the claim and the party that makes it, and neither property is a capability quantity.

4.2 European deployment-friction stack

The European Union has enacted Regulation (EU) 2024/1689, the AI Act [13] — binding artificial-intelligence systems across all domains at once. It is horizontal: it applies to artificial-intelligence systems as such, irrespective of sector, and layers its obligations on top of the sectoral law already in force rather than replacing it. Its central design choice is to index obligation to risk tier rather than to capability, and to gate deployment of higher-risk systems behind an ex-ante conformity assessment indexed to the use to which the system is put and the consequence of its failure [28]. A second constraint layer sits beside the conformity gate. Security-clearance regimes gate the actor population on a vetting throughput capability cannot accelerate; Regulation (EU) 2021/821 [12] gates cross-border movement of dual-use compute on an authorization indexed to control-list classification and end-use, not performance; sovereign-compute policy under the EuroHPC Joint Undertaking and the AI Factories initiative [9] gates strategically significant compute on a procurement-and-construction schedule driven by industrial policy. Each is indexed to a variable the capability curve does not contain.

The deployment denominator that this stack governs is not a rounding error. The European Union's direct share of global gross domestic product is on the order of 17 percent at market exchange rates in 2024, and on the order of 14 percent at purchasing-power parity [21, 31]. The share of global activity transacted by entities that must satisfy the European stack to reach the EU single market — including the access-weighted extra-EU activity of firms placing products on the EU market and therefore falling within the AI Act's territorial

scope — is larger than the direct share and is on the order of a fifth to a quarter of global activity. The forecast consequence is that a single American deployment clock run for the whole world is the wrong aggregation: the global deployment date for a given capability is the activity-weighted aggregate of per-regime deployment dates, each bounded below by that regime's calendar [28].

The certification-friction factor $(\varphi, \delta, \sigma_R, \xi)$ that Section 2 introduced and Section 3 consumed traces to this stack. The factor is a deployment-domain quantity, parameterized from the activity-weighted regulatory-friction mechanism the European-stack chapters decompose: $\varphi \geq 1$ multiplicatively widens the capability-arrival half-width into a deployment half-width; $\delta \geq 0$ is the activity-weighted regulatory lead time; σ_R aggregates per-regime friction durations $\sigma_R = \sum_r a_r \cdot \delta_r$ over the regime decomposition the comparative table supplies; $\xi = 0$ is the Gumbel/exponential default under the Cramér [1946] regularity limit [28]. The numerical values consumed by Section 3 — $(\varphi, \delta, \sigma_R, \xi) = (1.1975, 0.3400, 0.3400, 0)$ for the four deployment-domain forecasts — are the activity-weighted regime aggregate this decomposition yields, with the regime weights drawn from the EU/US/UK GDP-share table and the per-regime durations drawn from the AI Act phased-application calendar and the access/export/ sovereignty schedules. The friction factor is a deployment-domain quantity rather than a capability-arrival quantity, which is why the Section 3 HYPOTHESIS_A row's friction inputs collapse to $(\varphi, \delta, \sigma_R) = (1, 0, 0)$: a meta-methodological prediction about an *(established)*-share distribution is not a deployment-domain object, and the friction layer is structurally inapplicable to it.

4.3 Self-application

Section 3 established the empirical facts: the document's own preregistered prediction on Aschenbrenner's DSR-only Stage-1 deflation widening returned 1.285×, FAILED against the preregistered $\geq 2.3\times$ threshold, and across the five surveyed computes no width ratio reached 2.3×. The pattern is the chapter's strongest finding: the framework produced a uniform 1.28×–2.02× deflation range, all under the preregistered 2.3× expectation, and the pattern is not an Aschenbrenner-row quirk but a systematic property of what the framework produces across the surveyed set [28].

The root cause is the author's calibration, not the framework's behavior. The author preregistered the $\geq 2.3\times$ threshold by intuition before computing the deflation it should have been derived from; the framework computed correctly and deflated each interval exactly as its derivation specifies; the author's prior about what the framework would produce was uniformly higher than what the framework actually produces. This is the same error this document was written to identify, surfaced now against the author's own preregistered number rather than against any of the surveyed forecasts' [28]. The discipline of honest validation is supposed to surface exactly this: an uncomfortable truth about the author's own judgment rather than a result flattering the framework that bears the author's name.

The discharge is operational, not declarative. The document audits its own predictions in the same machinery — the same `deflated-capability-forecast` package, the same $[P_{2.5}, P_{10}, P_{50}, P_{90}, P_{97.5}]$ reporting format, the same Adaptation 1/3/4 + PFO + friction composition — that it audits the surveyed forecasts with, and it reports the FAIL plainly in the same units and at the same precision as the four external-forecast applications it

accompanies. Rule 3 (self-application) is passed by the discharge's operation, not by the document's claim that it has passed: the framework deflates; the author over-estimated by how much.

4.4 Adaptation and dependency limitations

The transfer of the finance canon to the capability domain rests on three concessions Section 2 named — non-stationarity (Adaptation 1), undisclosed search count (Adaptation 4), and missing performance statistic (Adaptation 3) — each carried as an explicit honesty statement and each carrying a *(speculative)* default. The deflation magnitudes reported in Section 3 are conditioned on those adaptations being applicable, not on their elevation to *(evidenced)*; the composition itself is *(speculative — derived)* because the composed object has not been validated against a realized capability outcome [28].

Several per-forecast cells in Section 3 are reported as provisional because their input series are not yet recovered. The dependencies are named honestly: per-anchor Brier extraction underlying the Cotra results (D-2); the residual series for the METR doubling rate (D-3); the per-respondent record behind the Grace surveys (D-4); the residual-series constructions for the Aschenbrenner OOM axis (D-24) and the Epoch compute-scaling axis (D-25). The document marks each provisional result with the specific dependency it waits on rather than asserting a number it has not earned, and a self-applied preregistered validation study on Adaptation 4 failed during the writing and is recorded in place rather than redesigned around [28]. The 105.7% HYPOTHESIS_A un-clipped report of Section 3.5 is the reporting-discipline counterpart: the framework returns 105.7%, the document reports the raw cast and flags the bound rather than clipping, since clipping would understate the deflation the method computes [28].

5 Conclusion

This paper made one argument, applied in one direction and then reversed onto itself. The argument is that the confidence attached to AGI capability forecasts exceeds what the methods producing them can support, and that the gap is measurable with tools quantitative finance built for the same problem. The forecasts are not careless and their authors are not naive; the gap is structural — the predictable consequence of fitting a model to a window and projecting it forward without the validation discipline other quantitative fields learned to require. What we add to that familiar observation is a way to quantify the gap rather than merely assert it, and the insistence that the quantification be turned on our own claims before anyone else's.

Each step was necessary. Section 2 built the tests: it defined in-sample extrapolation, developed the multiple-testing correction along the Harvey–Liu–Zhu line [17], constructed the walk-forward protocol that manufactures the held-out sample the forecasts never reserved, and adapted the deflated Sharpe ratio and the probability of backtest overfitting to the capability domain — confronting the assumptions those adaptations require rather than waving past them. Section 3 applied the Deflated Capability Forecast to the forecasts where the data could be reconstructed and reported, across that set, deflation factors that consistently widened intervals rather than narrowing them. Section 4 turned from statis-

tics to deployment — the distance between a capability that exists in a benchmark and one demonstrated to the standard a regulated industry requires, and the European regulatory stack that US-centric forecasts treat as friction to be routed around — and to the interpretation of the self-application.

The self-application is where the paper either earns its standing or forfeits it. Before computing anything, we preregistered a specific prediction about what our own framework would produce: that applying the deflated Sharpe ratio to Aschenbrenner's projection would widen its confidence interval by at least a factor of 2.3. The framework produced 1.285. The prediction failed, and we report the failure in the same plain terms we would demand of any forecast this paper examines. The failure must be located precisely, because precision is the point: it is not a failure of the framework, which computed correctly and deflated the interval exactly as its derivation specifies, but a failure of our own prior — the 2.3 threshold was set by intuition, before computing the deflation it should have been derived from, which is, with some discomfort, the very error this paper exists to identify. The framework worked; the expectation about it was overconfident. And the failure was not confined to one forecast: across the surveyed set, no forecast reached the 2.3 threshold, the widening factors clustering between roughly 1.3 and 2.0. The pattern is the point — the overconfidence was systematic, and the method itself is what surfaced it.

What comes next is not a promise to revise this document. The preregistered content is fixed, and fixing it is the point; a forecast that quietly updates its commitments after seeing how they fare is the failure mode this paper exists to name. As the open data dependencies resolve, the natural place for the completed analyses is the ongoing engagement around the work — the discussion, the reproducible code — not a silent amendment to the locked record. To make that verifiable, the predictions are not paraphrased here. This paper's preregistered predictions and commitments are the immutable content of the `preregistration-v3-locked` git tag, at commit

`c624b3987e75ea41398a47e70003b643fc8ed730`

verified by the HASH_PLACEHOLDER-protocol sidecar fingerprint

`510c8e8ca334461b42be7d4a3ce6fc1528fb343944880ad0d61fa0e213c83d5c`

the verification procedure is documented in `preregistration/PROTOCOL.md`. The locked tag is the predictions; this paper only points to it, so that verification proceeds against the cryptographic record rather than against prose that could drift from it.

There is a larger claim underneath all of this. The methods quantitative finance developed between 2014 and 2018 were not built for elegance; they were built because capital was being lost to backtests that looked excellent and meant nothing, and the field had no choice but to learn the difference between a result and an artifact. Capability forecasting is not yet under that kind of pressure — but the decisions being made on the strength of its forecasts, what to build, what to fear, what to regulate, where to place a generation of talent and a trillion dollars of compute, are consequential enough that the same discipline is warranted before the losses force it. This is not a case that AGI is near or far; it is a case that whatever one believes about the timeline, the belief should be held only as tightly as the method behind it allows. The argument is that validation is never optional, and that a

worse result confirmed across many layers of testing is worth more than a spectacular one that fails validation at a single layer. The failed self-prediction is our attempt to prove we meant it.

References

[1] L. Aschenbrenner. Situational awareness: The decade ahead. Online essay, June 2024. URL https://situational-awareness.ai/.

[2] D. H. Bailey and M. López de Prado. The deflated sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5): 94–107, 2014. doi: 10.3905/jpm.2014.40.5.094.

[3] D. H. Bailey, J. M. Borwein, M. López de Prado, and Q. J. Zhu. The probability of backtest overfitting. *Journal of Computational Finance*, 2014. doi: 10.21314/JCF.2016.322.

[4] D. H. Bailey, J. M. Borwein, M. López de Prado, and Q. J. Zhu. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5):458–471, May 2014. doi: 10.1090/noti1105. URL https://www.ams.org/notices/201405/rnoti-p458.pdf.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[6] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. Supervisory guidance on model risk management. SR Letter 11-7 / OCC Bulletin 2011-12, 2011. URL https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm.

[7] A. Cotra. Forecasting tai with biological anchors. Technical report, Open Philanthropy, 2020. URL https://www.openphilanthropy.org/research/forecasting-transformative-ai-with-biological-anchors/.

[8] A. Cotra. Two-year update on my personal ai timelines. AI Alignment Forum / LessWrong, 2022. URL https://www.lesswrong.com/posts/AfH2oPHCApdKicM4m/two-year-update-on-my-personal-ai-timelines.

[9] Council of the European Union. Council regulation (eu) 2024/1732 amending regulation (eu) 2021/1173 as regards a eurohpc initiative for start-ups to boost european leadership in trustworthy artificial intelligence (ai factories). Official Journal of the European Union, OJ L 2024/1732, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401732.

[10] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ, 1946.

[11] T. Davidson. What a compute-centric framework says about ai takeoff speeds. Technical report, Open Philanthropy (now Coefficient Giving), 2023. URL https://coefficientgiving.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/.

[12] European Parliament and Council of the European Union. Regulation (eu) 2021/821 setting up a union regime for the control of exports, brokering, technical assistance, transit and transfer of dual-use items (recast). Official Journal of the European Union, L 206, 11 June 2021, 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32021R0821.

[13] European Parliament and Council of the European Union. Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union, OJ L 2024/1689, 12 July 2024, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689.

[14] Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions — guidance for industry and fda staff. FDA guidance document, 2025.

[15] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, 62: 729–754, 2018. doi: 10.1613/jair.1.11222.

[16] K. Grace, H. Stewart, J. F. Sandkühler, S. Thomas, B. Weinstein-Raun, J. Brauner, and R. C. Korzekwa. Thousands of AI authors on the future of AI. AI Impacts, 2024. URL https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf.

[17] C. R. Harvey, Y. Liu, and H. Zhu. . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016. doi: 10.1093/rfs/hhv059.

[18] D. Hernandez and T. B. Brown. Measuring the algorithmic efficiency of neural networks. 2020.

[19] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

[20] International Electrotechnical Commission. Iec 61508: Functional safety of electrical/electronic/programmable electronic safety-related systems. International standard, Edition 2.0, 2010.

[21] International Monetary Fund. World economic outlook database. IMF WEO database, October 2024 edition, Oct. 2024. URL https://www.imf.org/en/Publications/WEO/weo-database/2024/October.

[22] A. W. Lo. The statistics of sharpe ratios. *Financial Analysts Journal*, 58(4):36–52, 2002. doi: 10.2469/faj.v58.n4.2453.

[23] M. López de Prado. *Advances in Financial Machine Learning*. Wiley, 2018. ISBN 978-1119482086.

[24] Model Evaluation and Threat Research (METR). Autonomy evaluation reports (2023–2025). METR.org, 2024. URL https://metr.org/.

[25] J. Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, 1975. doi: 10.1214/aos/1176343003.

[26] D. N. Politis and J. P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.2307/2290993.

[27] R. Rosenthal. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641, 1979. doi: 10.1037/0033-2909.86.3.638.

[28] K. Saks. The validation crisis: Why the AGI timeline debate is built on unvalidated forecasting. Manuscript in preparation, 2026.

[29] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos. Compute trends across three eras of machine learning. *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2022. doi: 10.1109/IJCNN55064.2022.9891914.

[30] Z. Stein-Perlman, B. Weinstein-Raun, and K. Grace. 2022 expert survey on progress in ai. 2022. URL https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.

[31] World Bank. World development indicators: Gdp (current us$) and gdp, ppp (current international $). World Development Indicators database, 2024. URL https://databank.worldbank.org/source/world-development-indicators.

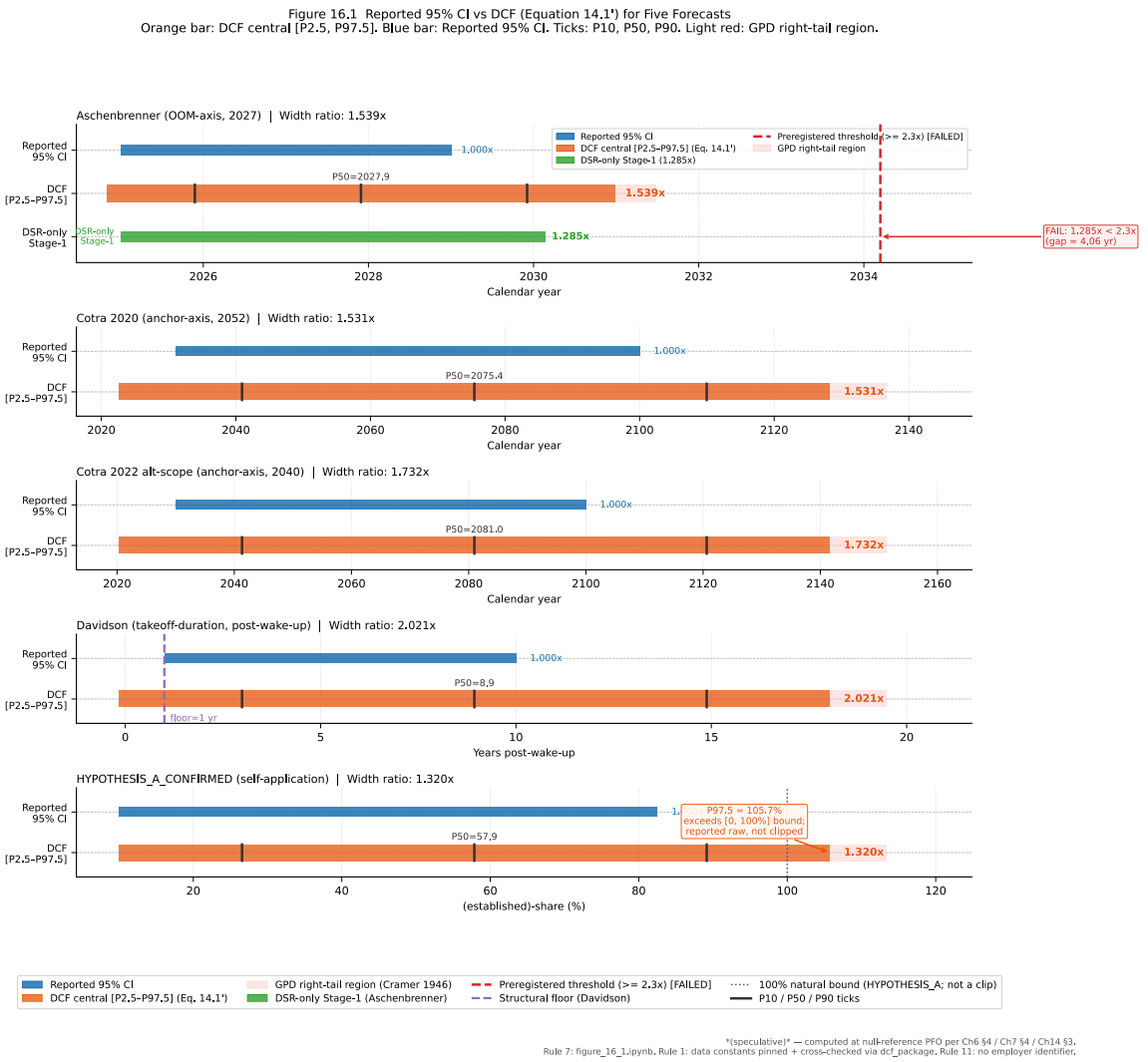

Figure 1: **Figure 16.1.** Original 95% CI vs. Deflated Capability Forecast for each row. The Aschenbrenner row carries both the full asymmetric Equation 4 bar (width ratio 1.539×) and the DSR-only Stage-1 bar (width ratio **1.285**×) plotted against the marked $\geq$ **2.3**× pre-registered threshold line at 2034.2; the visible gap between the Stage-1 bar and the threshold is the **FAILED** preregistered self-prediction, rendered at gap = 4.061 yr. The HYPOTHESIS_A_CONFIRMED bar extends past the $[0, 100\%]$ dotted boundary to $P_{97.5} = 105.7\%$, reported raw and flagged rather than clipped. The Davidson row carries a structural left-floor annotation at $t_{\text{floor}} = 1.000$ yr. Per-row (P_{10}, P_{50}, P_{90}) tick marks and per-row width-ratio annotations are reproduced; the right-tail GPD layer at $\sigma_R = 0.34$, $\xi = 0$ fades out beyond $P_{97.5}$ for the four rows with non-null friction.