ONE CAPABILITY OR MANY? STRUCTURAL AND PREDICTIVE TESTS OF BENCHMARK VALIDITY DISAGREE ABOUT ECONOMIC BENCHMARKS FOR FRONTIER AI

Anonymous authors
Paper under double-blind review

ABSTRACT

Frontier-model leaderboards now rank systems on economic benchmarks, and those rankings inform what organisations buy and what regulators scrutinise. Whether such benchmarks measure a capability distinct from general test-taking is a question of construct validity that a structural test and a predictive test can answer in opposite ways. We show that they do on a hash-pinned snapshot of a frontier leaderboard with 421 model configurations across twelve benchmarks, four of them economic, of which 103 configurations carry all three sparsely scored economic benchmarks and 96 carry all twelve; four hypotheses and their thresholds were fixed before analysis, and every deviation from the plan is reported. The first factor of a three-factor extraction carries 74.5% of common variance and tracks release date ($R^2 = 0.505$), and date adjustment lowers its share by 14.9 points. Under the dimensionality rule fixed in advance the economic benchmarks form no factor of their own. A leave-one-benchmark-out test with factors re-estimated inside every fold nevertheless finds that a multi-factor representation predicts held-out economic scores better than a single general index (pooled ΔMSE 0.037, 95% bootstrap interval $[0.019, 0.055]$) under the linear learners that fit best, an advantage that reverses for tree learners. Under the linear learners the same representation also predicts the eight other benchmarks better, so the battery carries predictive structure that one index misses and the economic benchmarks share it without forming a distinct factor. Construct validity should therefore be assessed by predictive tests alongside structural ones. We give a two-test protocol for benchmark builders and release the pinned data, the analysis plan and the code.

1 INTRODUCTION

Public leaderboards have become the dominant instrument for comparing and procuring frontier language models, and alongside academic knowledge tests they now carry *economic* benchmarks, suites built to proxy paid professional work and multi-step tool use. Whether those benchmarks measure a capability distinct from general test-taking is a question of construct validity, the degree to which a measurement captures the construct it claims to (Cronbach & Meehl, 1955; Messick, 1998), and the two standard tests of construct validity, one structural and one predictive, can disagree about the same benchmarks. The structural test asks whether a group of benchmarks forms a factor of its own once the variance they share with every other benchmark is accounted for, the discriminant criterion of Campbell & Fiske (1959). The predictive test asks whether the rest of the battery predicts them better out of sample through a richer representation than through a single general index, an adaptation of the incremental-validity criterion of Sechrest (1963) to a battery without an external criterion. On a snapshot of one frontier leaderboard the structural test says no and the predictive test says yes. Under a dimensionality rule fixed before analysis the economic benchmarks form no distinct factor, yet under the linear learners that fit best a multi-factor representation predicts held-out economic scores better than a general index, by a modest margin with an interval that excludes zero. The advantage extends to the other eight benchmarks, so the disagreement concerns how many dimensions the battery carries. This paper is about that disagreement, what produces it, and what it implies for how the validity of economic benchmarks should be judged.

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

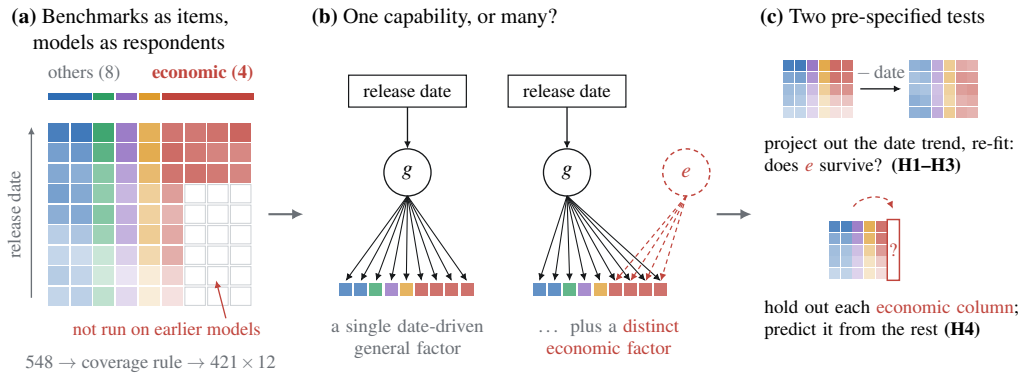


Figure 1: **Study design.** (a) Benchmarks as items (columns, economic in red) and model configurations as respondents (rows, ordered by release date); a coverage rule from the analysis plan reduces 548 configurations to a 421×12 matrix. (b) The two candidate structures, g alone or g with a distinct economic factor e. (c) The two pre-specified tests, date adjustment then re-fitting (H1–H3) and leave-one-benchmark-out prediction (H4); grids are stylised.

The question has stakes because of who reads leaderboards and what they now list. Alongside GPQA (Graduate-Level Google-Proof Q&A) (Rein et al., 2024) and Humanity's Last Exam (HLE) (Phan et al., 2025), the leaderboard studied here carries GDPval (Patwardhan et al., 2026), the τ-Bench family (Yao et al., 2025; Barres et al., 2026), and terminal or agentic coding suites (Merrill et al., 2026), and rankings on these columns inform what organisations buy and what regulators scrutinise. A separate column is warranted only if they measure something a single "intelligence" score does not. Machine learning has been urged to hold its benchmarks to that standard (Raji et al., 2021; Jacobs & Wallach, 2021; Bowman & Dahl, 2021), audits find construct validity under-examined in AI evaluation (Bean et al., 2025; Kearns, 2026). What remains open is whether an economic benchmark adds anything a general index lacks when the held-out column never enters estimation, and whether the structural and predictive answers to that question agree. The question is open because the nearest study compares individual benchmarks pairwise over ten models (Section 2), and the data and methods to answer it at leaderboard scale exist.

We frame a snapshot of frontier models scored across twelve benchmarks as a psychometric problem, with benchmarks as items, models as respondents, and scores as the responses from which latent structure is inferred, and answer it under a pre-specified design. Figure 1 shows the score matrix (a), the two candidate structures (b) and the two tests (c). Four research questions organise the analysis, each mapping to one pre-specified hypothesis (Section 4.3); the first three concern structure (Task 1, unsupervised) and the fourth prediction (Task 2, supervised).

- **RQ1 (dominance)** How many capabilities does the battery measure, and how dominant is the first?
- **RQ2 (release date)** How much of the leading axis is a release-date trend, and how much survives date adjustment?
- **RQ3 (structural distinctiveness)** After date adjustment, do the economic benchmarks load on a factor of their own?
- **RQ4 (predictive validity)** Does a multi-factor representation beat a single general index on held-out economic scores?

The paper makes four contributions. First, it shows that a structural and a predictive test of construct validity return opposite verdicts on the same economic benchmarks, and it argues from that dissociation that validity claims for a benchmark should say which test they rest on. Second, it introduces a leakage-safe leave-one-benchmark-out predictive test against a general-index baseline and finds that, under linear learners, the multi-factor advantage holds across the battery. Third, it measures how much of the dominant axis tracks model release date, which on a frontier snapshot carries what scale carries across generations. Fourth, it delivers the analysis under pre-specified hypotheses and thresholds with hash-pinned inputs, documenting every deviation and reporting the two failed tests as

failed. Section 2 places the study in the literature, Sections 3 and 4 give the data and the two tasks, Section 5 the results by hypothesis, and Section 6 what the disagreement implies.

2 Related work

Two strands of the construct-validity programme bear on this paper. The first treats LLM scores as psychometric data, following the case for measuring artificial systems with the instruments built for measuring minds (Hernández-Orallo, 2017) and the century-old observation that cognitive tests yield a positive manifold and a general factor (Spearman, 1904). Ilić & Gignac (2024) recover that manifold across 591 models on twelve tests, with a mean inter-test correlation of 0.73 and a general factor correlating with parameter count at about 0.6. How many further factors accompany the general one depends on the retention rule as much as on the data. Burnell et al. (2023), on 29 LLMs and 27 tasks, resolve capability into three factors; Maimon et al. (2026) recover eight from 60 models on 44 benchmarks under Kaiser's rule and a cumulative-variance threshold. That sensitivity is one reason to pair a structural test with a predictive one. Ruan et al. (2024) use principal components of a score matrix as a capability space in which performance, agentic performance included, becomes predictable (see also Zhou et al., 2025). Ho et al. (2025) stitch 38 benchmarks onto a single capability scale over 179 models, which is the one-dimensional assumption our single-index baseline embodies. Closest to our design, Kearns (2026) makes the validity question quantitative on the Open LLM Leaderboard, where a plain factor analysis fits a single factor explaining 72% of communal variance that tracks parameter count. Residualising every score on a fitted scaling law then lowers the mean inter-item correlation from 0.64 to 0.48, and that residualisation is the template for our date adjustment. Our design departs from his in three respects. The population is a frontier snapshot on which parameter counts are mostly unpublished, so release date takes the role scale plays there; the economic benchmarks are isolated as a block with a pre-specified discriminant test; and a leave-one-benchmark-out predictive test sits beside the structural one, which his study does not run. Across all of these studies scale is the variable to control; on a frontier snapshot we find that release date takes that role and compute adds little once date is removed, which suggests the variable to adjust for has moved with the field.

The second strand audits benchmarks as artefacts. Calls to repair benchmarking predate the frontier era (Bowman & Dahl, 2021; Kiela et al., 2021), and holistic evaluation frameworks (Liang et al., 2023), reproducibility guidance (Biderman et al., 2024), and proposals for a full evaluation science (Weidinger et al., 2025) have since sharpened them. Reviewing 445 benchmarks, Bean et al. (2025) find that 53.4% present any evidence for the construct validity of their benchmark and 16.0% report uncertainty estimates or statistical tests when comparing models. On the temporal side, Akhtar et al. (2026) find that nearly half of sixty benchmarks are saturated and that saturation rises with benchmark age, which makes release date a first-order variable whenever capabilities are compared across a moving frontier. Closest in question, Wan et al. (2026) introduce occupation-level benchmarks and test whether they predict GDPval scores better than generic capability benchmarks do, comparing Spearman correlations benchmark by benchmark across ten models, and find that their benchmarks win or tie in most comparisons. Our test differs in what it holds fixed. The baseline is a general index built from the whole battery where EconEvals compares one benchmark against another; the factors are re-estimated inside every fold so the held-out column never enters estimation; the thresholds were fixed before analysis; and the sample is a hash-pinned snapshot of 421 configurations. None of this work pairs a structural test with an out-of-sample predictive one on the same data, and none asks whether the two agree.

3 Data and design

Every number below rests on one hash-pinned snapshot and one coverage rule fixed before any score was inspected. The primary data set is a snapshot of the Artificial Analysis model leaderboard (Artificial Analysis, 2026), captured on 6 July 2026 directly from the public leaderboard page and pinned by its SHA-256 hash (Appendix N). The leaderboard is run by one independent operator that evaluates every model under a single harness on a fixed battery and publishes per-benchmark scores beside an aggregate index, so scores are comparable across providers. The snapshot contains 548 model configurations with per-benchmark scores, release dates, and provider metadata, where a configuration is one base model under one reasoning or effort setting. A secondary source, the Epoch

AI Notable AI Models data set (Epoch AI, 2026), supplies training-compute figures for a subset of models and serves only a scale robustness check. Both snapshots are pinned, so every number here is reproducible from fixed inputs.

To avoid selecting benchmarks or models on the very outcomes under study, we fixed the coverage rule in advance. A benchmark is retained when at least sixty models carry a score, and a model when it carries at least eight of the thirteen sufficiently covered benchmarks. Both floors were set as indicative values in the analysis plan and confirmed unchanged at the coverage audit; that confirmation used coverage counts alone, and no score, correlation or factor solution entered it. Appendix G shows what each protects. Under this rule thirteen of fourteen candidates pass; we drop APEX-Agents (Vidgen et al., 2026) for sparsity and demote MMMU-Pro (Yue et al., 2025) to a sensitivity-only role, since it is the battery's only multimodal benchmark and requiring it would cut the complete-case grid from 96 to at most 66 models (Appendix G), which leaves twelve primary benchmarks and 421 model configurations. Four of the twelve are economic benchmarks proxying paid professional or agentic work, and eight span academic knowledge, scientific coding (Tian et al., 2024; Zhu et al., 2025), long-context retrieval, and instruction following (Pyatkin et al., 2025) (Appendix F). The 421 rows are configurations, since one base model can appear under several reasoning or effort settings; the plan deduplicates to one row per base model as robustness check R1, and we keep the configuration level as primary because it is the more conservative choice.

Because the economic benchmarks are scored on far fewer models than the academic core, several nested samples appear in what follows, and each is named with its denominator. Of the 548 configurations parsed, 421 pass the coverage rule; 103 of those carry all three sparsely scored economic benchmarks (GDPval, Terminal-Bench v2.1, and τ^3-Banking), the count on which the coverage audit gates; and 96 carry all twelve benchmarks. That complete-case grid of 96 is the only sample on which the economic block is complete, and it carries every hypothesis and the predictive test. The dense grid of 409 configurations scored on the nine near-universal benchmarks serves only the corroborating clustering (Appendix K), where sample size matters more than economic completeness. The 96-model grid is the set of frontier configurations on which the economic benchmarks have actually been run, and the claims below extend to that set and to no wider population.

Since the benchmarks are reported on incommensurable scales, mostly accuracy fractions alongside one hallucination-penalised score and GDPval's Elo rating (Appendix F), we standardise each benchmark to zero mean and unit variance before any analysis. In the supervised task the standardisation is computed on each training fold and applied to its held-out fold, so no held-out information enters the fit. Missingness is structural, so we use complete-case analysis and impute nothing. On the complete-case grid the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy is 0.933, well above the 0.8 threshold (Kaiser, 1974), and Bartlett's test of sphericity is decisively rejected (Bartlett, 1950), so factor analysis is appropriate. The battery is also highly collinear before any modelling, with a mean off-diagonal Spearman correlation of $\rho = 0.79$ and every benchmark pair positive, and the economic benchmarks correlate with the academic core almost as strongly as with one another (Appendix H). The manifold is therefore at least as strong here as the 0.73 that Ilić & Gignac (2024) report on a broader population. The scope is internal validity, the correlational and predictive structure of aggregate scores on one cross-section; item-level responses, external economic impact, and saturation dynamics lie outside it.

4 Methods

4.1 Task 1: latent structure

Each structural question has its own quantity, and Appendix D states each method formally. The first factor's share of common variance measures dominance (RQ1), the change in that share after date adjustment measures the release-date component (RQ2), and the date-adjusted loading pattern tests distinctiveness (RQ3). To fix dimensionality we use Horn's parallel analysis (Horn, 1965), which retains factors whose eigenvalue exceeds the 95th percentile of eigenvalues from random data of the same shape, with the scree criterion (Cattell, 1966) as a cross-check. On collinear data this rule is more conservative than the Kaiser rule (Kaiser, 1960) and than information criteria (Appendix D), and both rules retain one factor here. We prefer factor analysis to the principal-component description of Ruan et al. (2024) because components mix common and unique variance, whereas the common-factor model isolates shared capability, which is the distinction a validity question turns on. The model is

accordingly a maximum-likelihood exploratory factor analysis (EFA) with an oblique oblimin rotation (Jennrich & Sampson, 1966; Thurstone, 1947), chosen because the date-adjusted factors are strongly correlated (Section 5.3) and an orthogonal rotation would spread shared variance across artificially independent axes. Every factor is oriented so that higher means more capable. The estimators used throughout, this factor model, Thurstone factor scores, and the leave-one-benchmark-out prediction of Section 4.2, are implemented in a released and versioned library (Anonymous, 2026), whose acceptance tests lock the values reported below against the archived tables (Appendix N).

To separate a general capability from a shared time trend, we regress each standardised benchmark on release date, and on a subsample also on training compute, and re-fit the factor analysis on the residuals; we call this date adjustment. The adjustment does not treat time as noise. Later models arrive with newer data, algorithms and post-training, so the date-aligned component contains real progress, and the adjustment asks what the benchmarks measure beyond what the calendar already indexes. It removes from the benchmark covariance the rank-one component aligned with date (Appendix D), so the resulting change in the first factor's variance share measures how much of the apparent general factor the calendar already indexes, the analogue of the scaling-law residualisation Kearns (2026) reports. Clustering corroborates the structure without assuming a factor model (Appendix K).

4.2 Task 2: predictive validity

Task 2 recasts the distinctiveness question as a forecast, asking whether a model's score on a held-out economic benchmark follows from its behaviour on the others, a sharper test than the in-sample capability space of Ruan et al. (2024) because the target column is withheld from estimation entirely. To that end, the leave-one-benchmark-out (LOBO) protocol takes each economic benchmark as the target in turn and predicts it from representations of the remaining eleven on the $n = 96$ complete-case grid, so that success demands transfer across benchmarks. Rung (i) of a five-rung predictor ladder uses release timing and scale only, rung (ii) a single mean-score general index, rung (iii) the first factor only, rung (iv) the k factors (the three-factor solution of Section 5.3), and rung (v) the k factors plus model covariates. The contrast of interest sets rung (iv) against the single-index rung (ii).

To see whether the ladder comparison depends on model class, we fit each rung with four learners spanning the linear-regularised and tree families (Hastie et al., 2009), ridge (Hoerl & Kennard, 1970), elastic net (Zou & Hastie, 2005), random forest (Breiman, 2001), and gradient boosting (Friedman, 2001). Hyperparameters are tuned by grid search in an inner five-fold cross-validation nested inside an outer five-fold split that supplies the reported out-of-fold error, the nesting that removes selection bias from the estimate (Varma & Simon, 2006) (grid in Appendix I). The plan does not fix the learner for the H4 contrast. We report it ridge against ridge, the learner with the lowest out-of-fold error on both rungs, and Appendix L gives it under each of the four learners; the best learner per rung in Table 1 is chosen on out-of-fold error and is descriptive. The remaining eleven include the other economic benchmarks, so the test measures transfer within the battery from companions and academic columns alike; withholding the whole economic block at once is the sharper test we did not run. Because the factors are estimated quantities, we re-estimate them inside the loop by fitting the factor analysis on each training fold alone, projecting the held-out fold through the training-fold Thurstone factor-score weights, and re-orienting by the training-fold mean index. A naive fit-on-all-data pipeline would silently open that leakage path (full procedure in Appendix B).

Two error metrics then serve two purposes. Root mean squared error (RMSE) on the standardised target is the descriptive metric since it stays on the target's own scale. The test statistic is $\Delta\mathrm{MSE} = \mathrm{MSE}(\text{single index}) - \mathrm{MSE}(k \text{ factors})$ per economic target, positive when the multi-factor representation wins, with a 95% interval from a paired bootstrap of the pooled out-of-fold errors (Efron & Tibshirani, 1993) with $B = 2000$.

4.3 Pre-specified hypotheses

Before analysis, the design fixed four hypotheses with quantitative thresholds set with reference to the figures Kearns (2026) reports for the Open LLM Leaderboard. **H1** (dominance, RQ1) holds that the first factor carries more than 50% of common variance. **H2** (date artefact in the plan's term, which we read as association, RQ2) has two parts, (i) that the dominant factor's release-date R^2 is at least 0.30 and the strongest of the factors, and (ii) that date adjustment lowers the first factor's share by at

least fifteen percentage points. **H3** (structural distinctiveness, RQ3) holds that after date adjustment parallel analysis retains a factor on which the four economic benchmarks load at least 0.40 and exceed their cross-loadings, which is the discriminant criterion of Campbell & Fiske (1959) applied to a benchmark battery. **H4** (predictive validity, RQ4) holds that the multi-factor representation's ΔMSE interval excludes zero and a majority of the four economic targets individually favour it. We report every outcome, including the two that fail.

Two factor counts are used in what follows. The primary count is $k = 1$, returned by parallel analysis, the Kaiser rule, and the scree elbow, and it governs H1 to H3; the share statistic in H1 and H2(ii) is computed on the three-factor extraction, since under $k = 1$ it is 100% by construction (Appendix D). The second is $k = 3$, an exploratory over-extraction that no criterion selects, which supplies the loading pattern of Section 5.3 and the k-factor predictor of Task 2 and is a documented deviation from the plan (Appendix J). The predictive conclusion does not hinge on it, since sweeping rung (iv) over $k \in \{2, 3, 4, 5\}$ beats the single index at every count (Appendix L).

5 Results

The results follow the four hypotheses in order, and the two verdicts that fail are reported as they fell.

5.1 Dominance of a single factor (H1)

Every dimensionality diagnostic points to a dominant capability axis. The first principal component alone accounts for 79.4% of total score variance, and parallel analysis retains a single factor, since only the first observed eigenvalue of 9.53 exceeds its random-data 95th percentile of 1.79 (Appendix Figure 5). In the three-factor maximum-likelihood extraction the first factor holds 74.5% of common variance, above the 50% dominance threshold, so the dominant general factor reported across the psychometric literature (Section 2), including the 72% Kearns (2026) reports under a different specification, replicates on a frontier snapshot. H1 is therefore supported, which answers RQ1 in full, and because rank-based and logit-transformed variants only raise the share (Section 5.5), the Pearson-based headline is the conservative one.

5.2 Release date and the dominant factor (H2)

The factor's scores track model release date with a logistic-fit R^2 of 0.505 (0.477 under an ordinary-least-squares fit, Appendix L), which is the strongest date association of the three factors. H2(i) is therefore supported, and the leading axis of "capability" is closely associated with release date. Adjusting every benchmark for release date and re-fitting then lowers the first factor's common-variance share from 74.5% to 59.6% (Figure 2c), a drop of 14.9 percentage points. That point estimate falls just short of the pre-specified fifteen-point threshold, with a bootstrap 95% interval of $[-5.3, +32.7]$, so on the primary grid the drop is indistinguishable both from zero and from the threshold. We report H2(ii) as not met on the point estimate without moving the line after seeing the result.

Whether H2(ii) passes depends on the unit of analysis. Under the plan's base-model unit the drop is 24.1 points and clears the threshold; under the configuration-level primary it is 14.9 points and fails; on the 58-model compute-known subsample date alone gives 16.5 points and clears it. Two of the three specifications therefore pass. Promoting the deduplicated grid to primary would pass the threshold on fewer models (89 against 96) and turn the plan's robustness unit into the headline, so we keep the configuration level as primary. The drop clears the threshold on the subsample and the deduplicated grid, which we report as consistent with H2(ii) while the primary verdict stands as not met. RQ2 is therefore answered in part, since the leading axis tracks release date closely (H2(i)) while the pre-specified size of the correction is not met on the primary grid (H2(ii)). Adding training compute does not deepen the correction, since on the compute-known subsample the joint date-plus-compute drop is smaller at 9.3 points. The likely reason is that later models are also larger and the two predictors split their shared variance.

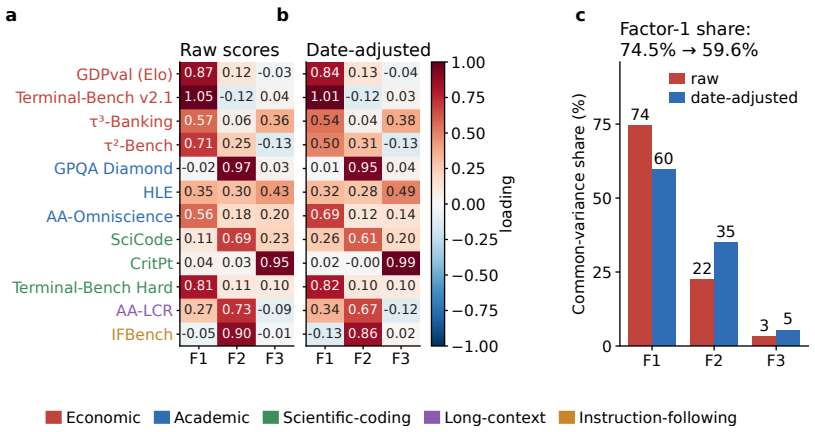


Figure 2: **Factor loadings from the three-factor oblimin EFA.** (a) Raw scores; the economic benchmarks (red) already concentrate on F1. (b) After date adjustment they still concentrate on F1, with a largest cross-loading of 0.38. (c) The first factor's common-variance share falls from 74.5% to 59.6%, a 14.9-point drop.

5.3 Economic distinctiveness (H3)

After date adjustment we test whether the economic benchmarks form a distinct factor, which requires that parallel analysis retain a factor and that its loadings concentrate discriminantly on the economic block. On the date-adjusted data parallel analysis retains a single factor, which leaves no block to separate; the separation appears only once three factors are extracted, which the dimensionality rule does not endorse. Holding H3 to the plan as H2(ii) was held to its 15-point line, we report H3 as not supported as specified, which answers RQ3 negatively under the plan.

Under the three-factor over-extraction the economic block does separate, and we report the pattern as exploratory. All four economic benchmarks load at least 0.40 on a single factor, F1, with GDPval at 0.84, Terminal-Bench v2.1 at 1.01, τ^3-Banking at 0.54, and τ^2-Bench at 0.50, against a largest economic cross-loading of 0.38 (Figure 2b, tabulated in Appendix E). The mean absolute F1 loading is 0.72 for the economic benchmarks and 0.32 for the rest. Because the three factors are strongly correlated (0.67 between F1 and F3 and 0.72 between F2 and F3) the separation reads as a secondary refinement of one dominant axis. Bootstrap intervals on the economic loadings are wide (0.10 to 0.78 for τ^2-Bench and 0.44 to 0.99 for GDPval), excluding zero but overlapping one another, so the pattern is suggestive and loosely estimated.

5.4 Predictive validity (H4)

The final test asks whether a multi-factor representation of the other benchmarks predicts a held-out economic benchmark better than their mean. Across the predictor ladder a single mean-score index is already a strong baseline, at economic-block test RMSE 0.474 and R^2 0.771; the k-factor representation improves on it to RMSE 0.433 and R^2 0.808, and adding covariates gives no further gain (Table 1, Figure 3a). The first factor alone predicts poorly at RMSE 0.950. Bootstrapping ΔMSE over the pooled out-of-fold errors gives a pooled improvement of $+0.037$ with a 95% interval of $[+0.019, +0.055]$, and because that resample includes near-duplicate settings of the same base model, we re-run it on the deduplicated grid of 89 distinct models, where the improvement is $+0.038$ with $[+0.020, +0.056]$ (Appendix L). Three of the four economic benchmarks improve individually, GDPval by $+0.026$, Terminal-Bench v2.1 by $+0.028$, and τ^3-Banking by $+0.056$, while τ^2-Bench's interval includes zero (Table 1, Figure 3b). H4 is therefore supported under ridge, the best-fitting learner, on the rule of Section 4.3; the plan fixed no learner, and Section 5.5 reports the tree-learner reversal.

The 0.037 gain sits on top of a single index that already explains 77% of economic-block variance. It shows that the battery carries structure a single index misses. Whether an economic benchmark adds information the others lack is a separate question, which needs the economic block withheld as a

Table 1: **Predictor-ladder cross-validated errors (economic block) and the H4 bootstrap test.** Left, train and test RMSE on standardised targets averaged over the economic targets, out-of-fold R^2, and the best learner per rung (RF random forest, EN elastic net). Right, $\Delta\text{MSE} = \text{MSE}(\text{mean index}) - \text{MSE}(k \text{ factors})$, $B = 2000$; every interval except τ^2-Bench's excludes zero.

Rung	Learner	Train	Test	R^2
(i) timing / scale	RF	0.521	0.741	0.459
(ii) mean index	Ridge	0.463	0.474	0.771
(iii) first factor	EN	0.933	0.950	0.110
(iv) k factors	Ridge	0.410	0.433	0.808
(v) k factors + cov.	Ridge	0.391	0.438	0.800

Target	ΔMSE 95% CI
GDPval (Elo)	$+0.026$ $[+0.010, +0.044]$
Terminal-Bench v2.1	$+0.028$ $[+0.001, +0.056]$
τ^3-Banking	$+0.056$ $[+0.008, +0.103]$
τ^2-Bench	$+0.038$ $[-0.011, +0.086]$
Pooled economic	$\mathbf{+0.037}$ $[\mathbf{+0.019}, \mathbf{+0.055}]$

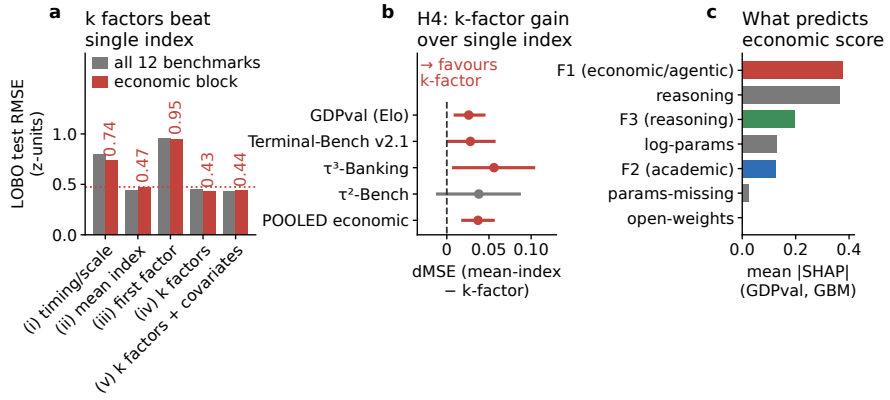


Figure 3: **Predictive validity under the LOBO protocol.** (a) Test RMSE by rung, economic block (red) and all twelve targets (grey). (b) Bootstrap ΔMSE (single index minus k factors) with 95% intervals; positive favours the multi-factor representation. (c) SHAP (SHapley Additive exPlanations) importances for the rung-(v) gradient-boosting model of GDPval, the learner that admits exact SHAP values.

whole or an external criterion. The one exception, τ^2-Bench, carries the widest interval, and it is also the economic benchmark with the lowest F1 loading (0.50) and the one that branches alone in the benchmark clustering (Appendix K), so its null is consistent with its structural separation from the other three.

5.5 Robustness

The headline structure is stable under perturbations to the unit of analysis, the correlation measure, the adjustment covariates, and the factor count (Appendices J and L). Collapsing the 421 configurations to one row per base model raises the first-factor share to 90.2% and the date-adjustment drop to 24.1 points, and rank-based and logit-transformed correlations raise the share to 94.5% and 91.8%, so H1 is no metric artefact (Schaeffer et al., 2023). The H4 verdict depends on neither the regularisation strength nor the factor count, and it does depend on learner family. Ridge and elastic net favour the multi-factor representation on all four targets, and the tree learners, which fit the economic block worse on both rungs, reverse it (C8). The k-factor advantage also extends beyond the economic block, averaging 0.048 ridge against ridge on the eight other targets (positive on seven) against 0.037 on the economic four, with intervals computed for the economic block only.

6 Discussion

Taken together, the two tasks answer the four research questions, and the answers do not line up. On RQ1 the battery is governed by one dominant axis holding 74.5% of common variance, consistent with prior reports; on RQ2 that axis tracks release date ($R^2 = 0.505$), and date adjustment lowers its share by about a fifth on the point estimate (74.5% to 59.6%), with an interval that spans zero on the

primary grid. On RQ3 the economic benchmarks form no separate factor under the pre-specified rule and separate only under an over-extraction, yet on RQ4 a multi-factor representation still predicts held-out economic scores better than a single general index, on three of four benchmarks and on the deduplicated grid, and it predicts the other benchmarks better too, all under the linear learners that fit best. The structural and predictive tests disagree, and the disagreement is the finding. A retention rule asks whether the economic benchmarks add enough shared variance to move an eigenvalue past a random-data threshold; a held-out prediction asks whether a richer representation of the battery predicts them better than a general index does. The second can be answered yes while the first is answered no, and two features of the study produce that gap without any flaw in either test. The tests set different bars, since a retention rule needs an eigenvalue above what random data of the same shape produce, where at $n = 96$ that line sits at 1.57 against an observed second eigenvalue of 0.91, while an out-of-sample gain of a few hundredths of a standardised unit clears a bootstrap interval. The sample is small for the structural test and adequate for the predictive one, which is why the bootstrap interval on the date drop spans zero while the interval on $\Delta\mathrm{MSE}$ does not. A single index reproduces most of what distinguishes models, 77% of economic-block variance out of sample, and that agreement does not license the claim that the economic benchmarks add nothing at the margin. A price index built from benchmark scores shows the same shape, since removing contamination-flagged benchmarks leaves its model ranking at a rank correlation of 0.998 while the index moves by 0.49 log points a year (Zhu, 2026b). Construct validity should therefore be assessed with a predictive test beside the structural one, and a validity claim for a benchmark should say which test it passed.

For leaderboard use, a single intelligence index captures most of what distinguishes models, and the structure it misses carries predictive information for every column, which justifies reporting the benchmarks separately, the economic ones among them. A leaderboard ranking models released months apart is ranking them largely along an axis that tracks release date. A comparison within one release window holds that trend fixed, as a buyer choosing between contemporaneous models needs.

A second contribution follows for benchmark builders in the form of a two-test protocol, since a benchmark's value for distinguishing contemporaneous models depends on the variance it adds once the shared date trend is removed. A builder who wants to show that a new suite measures something distinct can apply the two tests directly with what a leaderboard already publishes, per-model scores on the new benchmark and the existing battery and their release dates. Step one, the structural test, regresses every benchmark on release date, fits an oblique factor model to the residuals, and checks whether the new benchmark's loading separates from the general factor. Step two, the predictive test, holds the new benchmark out, predicts it from a k-factor representation of the others with factors re-estimated in every fold and a regularised linear learner, reporting tree learners beside it, and checks whether that beats a single mean-score index with a bootstrap interval excluding zero. A benchmark that passes both adds a construct, one that passes only the second shares structure with the battery beyond the general index without forming a factor of its own, and one that passes neither is re-measuring the progress trend. Step two presupposes companions in the battery, and a benchmark for a construct the battery does not yet represent needs an external criterion. An operator can run both tests on every new benchmark and publish the two statistics beside the score, and the accompanying code does so from a single score matrix.

Four limitations bound the claims, in decreasing order of consequence. First, the economic factor comes from an exploratory over-extraction, so it stands as a hypothesis for a confirmatory factor model on an independent snapshot. Second, the analysis is observational. Release date stands in for everything that arrives with newer models, so the component the adjustment removes contains real improvement along with any calendar effect, and the counterfactual that would separate them, an older and a newer model evaluated under the same conditions at the same time, is one a leaderboard cannot supply. Third, coverage is uneven, so the economic results carry wider uncertainty, and fourth, the results describe one leaderboard on one date. Both tests are internal to the leaderboard, so they leave open whether scores that pass them predict performance on naturalistic professional tasks with tools and interaction, assessed independently of the benchmarks and with any automated grader's validity measured (Zhu, 2026a).

Economic benchmarks therefore measure neither one capability nor many. Most of what they record is a general factor that tracks release date, and under the linear learners the structure one index misses predicts held-out scores without forming a factor of its own. Hypotheses and thresholds were fixed in advance, Appendix J lists every departure, and every verdict stands as it fell.

AI USE STATEMENT

Generative AI tools assisted with tasks ancillary to the research, namely reference verification and formatting, appendix and figure preparation, auditing and debugging of the released code, and consistency checks of numbers and cross-references against the archived outputs. They were not used to generate research questions or hypotheses, to design or run the analysis, or to draw its conclusions. The research and its text are our own, every AI-assisted output was reviewed and revised by us, and we take full responsibility for the final content.

ETHICS STATEMENT

The study analyses publicly reported benchmark scores of AI systems and involves no human subjects or personal data; the leaderboard snapshot is used under the operator's public-page terms and a secondary data set under a CC BY 4.0 licence. We declare no conflicts of interest or sponsorship relevant to the work.

REPRODUCIBILITY STATEMENT

Every number in the paper traces to two SHA-256-pinned snapshots described in Appendix N. The analysis code is released as a versioned library (Anonymous, 2026), extracted from the archived analysis notebook and released under the MIT licence, with the derived tables under CC BY 4.0 and the leaderboard snapshot redistributed under the operator's public-page terms; an anonymised repository and archive are provided in the supplementary material. One command recomputes the results reported here from the pinned 6 July 2026 snapshot, and every table the package emits states its reproducibility tier and its deviation from the archived value. Recomputed outputs are produced again from the pinned inputs and agree with the archived table within a stated tolerance; they cover the structure results (the KMO of 0.933, the first-factor share of 74.5%, the residualised share of 59.6% and its 14.9-point drop, and the residualised oblimin loadings) and the leave-one-benchmark-out point estimates, including the pooled economic $\Delta\mathrm{MSE}$ of $+0.037$ and the four-learner ladder, which reproduces the archived rung summary exactly, the summary's per-rung error being the mean of the per-target RMSEs. Statistically reproduced outputs are produced again under a random stream the archive does not record and agree within three Monte-Carlo standard deviations; they cover the bootstrap interval $[+0.019, +0.055]$ and the k-sweep. Regenerated outputs are copied from the archive and re-rendered without recomputation; they cover the clustering table, which lies outside the library's scope, and the logit first-factor share, the calibration constants, and the BIC row, whose computations do not appear in the archived notebook. Two quantities were reconstructed because the archive ships no code for them, and a reader should weigh them accordingly. The four-parameter logistic behind the date R^2 of 0.505 was recovered by fitting candidate forms until one reproduced the archived value; the analysis plan specifies a logistic fit without fixing its parameterisation, so the recovered form is consistent with the plan. The pooling rule behind the ladder table was inferred as the mean of per-target RMSEs, which reproduces all ten archived rows, whereas the notebook's own helper pools by root mean square and reproduces none. The analysis plan, fixed before the data were captured, is deposited in a public registry; the registry does not support anonymised links for completed deposits, so Appendix N gives the recorded timeline and the plan's checksum in place of one, and the deviations from the plan are documented in Appendix J.

REFERENCES

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, et al. When AI benchmarks plateau: A systematic study of benchmark saturation. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.16763.

Anonymous. A library for psychometric analysis of AI benchmark scores (v0.1.0). Computer software, 2026. Anonymised repository and archive in the supplementary material.

Artificial Analysis. Artificial Analysis LLM Leaderboard and Intelligence Index. https://artificialanalysis.ai, 2026. Snapshot accessed 6 July 2026.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2506.07982.

Maurice S. Bartlett. Tests of significance in factor analysis. *British Journal of Psychology (Statistical Section)*, 3:77–85, 1950.

Andrew M. Bean, Ryan O. Kearns, Angelika Romanou, Franziska S. Hafner, Harry Mayne, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, 2025. arXiv:2511.04703.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, et al. Lessons from the trenches on reproducible evaluation of language models, 2024. arXiv:2405.14782.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4843–4855, 2021.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Ryan Burnell, Han Hao, Andrew R. A. Conway, and José Hernández-Orallo. Revealing the structure of language model capabilities, 2023. arXiv:2306.10062.

Donald T. Campbell and Donald W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959.

Raymond B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1 (2):245–276, 1966.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

Epoch AI. Notable AI Models (data set, CC-BY 4.0). https://epoch.ai/data/notable-ai-models, 2026.

Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

José Hernández-Orallo. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press, 2017.

Anson Ho, Jean-Stanislas Denain, David Atanasov, Samuel Albanie, and Rohin Shah. A Rosetta stone for AI benchmarks, 2025. arXiv:2512.00193.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

John L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2): 179–185, 1965.

David Ilić and Gilles E. Gignac. Evidence of interrelated cognitive-like capabilities in large language models: indications of artificial general intelligence or achievement? *Intelligence*, 106:101858, 2024.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021.

Robert I. Jennrich and Paul F. Sampson. Rotation for simple loadings. *Psychometrika*, 31(3):313–323, 1966.

Henry F. Kaiser. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1):141–151, 1960.

Henry F. Kaiser. An index of factorial simplicity. *Psychometrika*, 39(1):31–36, 1974.

Ryan O. Kearns. Quantifying construct validity in large language model evaluations, 2026. MSc thesis, Oxford Internet Institute, University of Oxford, Trinity Term 2025. arXiv:2602.15532.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, et al. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4110–4124, 2021.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. arXiv:2211.09110.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pp. 4765–4774, 2017.

Aviya Maimon, Amir David Nisan Cohen, Gal Vishne, Shauli Ravfogel, and Reut Tsarfaty. From benchmarks to skills: Low-rank factors for LLM evaluation, 2026. arXiv:2507.20208.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces, 2026. arXiv:2601.11868.

Samuel Messick. Test validity: a matter of consequence. *Social Indicators Research*, 45:35–44, 1998.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, et al. GDPval: Evaluating AI model performance on real-world economically valuable tasks. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.04374.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, et al. Humanity's Last Exam, 2025. arXiv:2501.14249.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following, 2025. arXiv:2507.02833.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Datasets and Benchmarks Track*, 2021.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2311.12022.

Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Lee Sechrest. Incremental validity: A recommendation. *Educational and Psychological Measurement*, 23(1):153–158, 1963.

Quan Shi, Alexandra Zytek, Pedram Razavi, Karthik Narasimhan, and Victor Barres. τ-Knowledge: Evaluating conversational agents over unstructured knowledge, 2026. arXiv:2603.04370.

Charles Spearman. "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

Louis L. Thurstone. *Multiple-Factor Analysis*. University of Chicago Press, 1947.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, et al. SciCode: A research coding benchmark curated by scientists. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 2024. arXiv:2407.13168.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 63(2):411–423, 2001.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006.

Bertie Vidgen, Austin Mann, Abby Fennelly, John Wright Stanly, Lucas Rothman, Marco Burstein, et al. APEX-Agents, 2026. arXiv:2601.14242.

Alexander Wan, Stephane Hatgis-Kessell, Tomás Aguirre, Percy Liang, and Rishi Bommasani. Economic evaluations of language models, 2026. arXiv:2607.19375.

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, et al. Toward an evaluation science for generative AI systems, 2025. arXiv:2503.05336.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.12045.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, et al. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025. arXiv:2409.02813.

Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, et al. General scales unlock AI evaluation with explanatory and predictive power, 2025. arXiv:2503.06378.

Louis Yiven Zhu. Three ways classical test theory misleads for LLM judges, 2026a. arXiv:2609.29709.

Louis Yiven Zhu. The price of intelligence: A quality-adjusted price index for AI services, 2026b. arXiv:2608.29843.

Minhui Zhu, Minyang Tian, Xiaocheng Yang, Tianci Zhou, Lifan Yuan, Penghao Zhu, Eli Chertkov, Shengyan Liu, Yufeng Du, Ziming Ji, et al. Probing the critical point (CritPt) of AI reasoning: A frontier physics research benchmark, 2025. arXiv:2509.26574.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.

APPENDIX

The appendix collects the material that supports the main text without interrupting its argument. Each section opens with the main-text section it supports, and every figure and table is placed with the section that discusses it.

A ABBREVIATIONS AND NOTATION

Table 2: Abbreviations and symbols, in order of first appearance in the main text.

GPQA	Graduate-Level Google-Proof Q&A	LOBO	leave-one-benchmark-out
HLE	Humanity's Last Exam	RMSE	root mean squared error
LLM	large language model	MSE	mean squared error
RQ	research question	ΔMSE	$\mathrm{MSE}(\text{single index}) - \mathrm{MSE}(k \text{ factors})$
H1–H4	pre-specified hypotheses	CI	confidence interval
KMO	Kaiser–Meyer–Olkin sampling adequacy	OLS	ordinary least squares
EFA	exploratory factor analysis	BIC	Bayesian information criterion
PCA	principal-component analysis	SHAP	SHapley Additive exPlanations
ML	maximum likelihood	RF, EN	random forest, elastic net
R^2	coefficient of determination	GBM	gradient-boosting machine
ρ	Spearman rank correlation	R1, C1–C8	planned and delivered robustness checks
k	number of retained factors	AA	Artificial Analysis (leaderboard operator)
Λ, Φ, Ψ	loadings, factor covariance, uniquenesses	W	Thurstone factor-score weights

B LEAVE-ONE-BENCHMARK-OUT PROCEDURE

Supports Section 4.2, which describes the protocol in words. Figure 4 states the complete procedure, including the two safeguards on which the predictive test depends. First, the factor model is re-fitted inside every outer training fold, so the held-out models never influence the loadings used to score them. Second, hyperparameters are tuned in an inner loop nested within the outer split (Varma & Simon, 2006), and the H4 contrast is reported ridge against ridge; the plan does not fix that learner, and ridge is also the lowest-error learner on both rungs, so check C8 in Appendix L gives the contrast under all four. The best learner per rung in Table 1 is chosen on pooled out-of-fold error, as the procedure below shows, so that column is descriptive and its errors carry the optimism of that choice.

Procedure Leave-one-benchmark-out predictive validity with nested cross-validation
Input: raw score matrix X (models $\times$ benchmarks), economic targets T, learners L, rungs R.
for each target $t \in T$ **do**
  partition models into 5 outer folds
  for each outer fold (train, test) **do**
    drop column t; fit standardisation and EFA on train only; form Thurstone weights $W = R^{-1}\Lambda$
    project test models through W; orient factors by train mean index
    for each rung $r \in R$ and learner $\ell \in L$ **do**
      tune ℓ by inner 5-fold grid search on train; predict test, store out-of-fold error
  select best learner per rung by pooled out-of-fold RMSE (the H4 contrast is ridge against ridge and does not depend on this choice)
compute $\Delta\mathrm{MSE} = \mathrm{MSE}(\text{rung ii}) - \mathrm{MSE}(\text{rung iv})$; bootstrap $B = 2000$ for 95% CI
Output: per-target and pooled ΔMSE with confidence intervals (H4).

Figure 4: The leave-one-benchmark-out procedure with leakage-safe factor projection and nested tuning.

C POSITIONALITY, DIRECTIONALITY, AND THREATS TO VALIDITY

Supports Section 4.3. A factor solution is identified only up to sign, so we state the direction of every factor explicitly. We orient each factor to correlate positively with the model's mean standardised score. "Higher" then always means "more capable", and every loading in Table 3 reads in the same

direction. The choice is not cosmetic, since an un-oriented solution could report the same economic factor with flipped signs and invite the opposite interpretation. The direction of the date adjustment is equally deliberate. We regress scores on release date and analyse the residuals, which measures how much benchmark co-movement survives once the release date is known. We do not run the reverse regression, as we make no claim that the general factor causes the calendar. Finally, we fix the sign of the predictive contrast in advance, so that $\Delta\mathrm{MSE}$ is the single-index error minus the k-factor error, a positive value favours the richer representation, and H4 requires this positive interval to exclude zero.

Two threats to validity follow from the study's vantage point. First, we observe only the models that vendors chose to submit to a public leaderboard. The sample is therefore a convenience sample of frontier systems, and the claims generalise to models on this leaderboard alone. Second, the economic benchmarks are newer and sparser, so any economic-specific structure is estimated from fewer observations and is more sensitive to the particular models that carry those scores. The wide bootstrap intervals in Table 1 quantify that limitation.

D MATHEMATICAL BASIS OF THE TWO TASKS

Supports Sections 4.1 and 4.2. Each subsection states a component of the method, derives the quantity the main text reports, and gives the reason the component was chosen over its nearest alternative.

The common-factor model and the dominance statistic. Let $\mathbf{z} \in \mathbb{R}^p$ be the vector of p per-benchmark standardised scores for a model. The common-factor model writes

$$\mathbf{z} = \Lambda\mathbf{f} + \boldsymbol{\varepsilon}, \qquad \Sigma = \Lambda\Phi\Lambda^\top + \Psi,$$

with $\Lambda \in \mathbb{R}^{p\times k}$ the loading matrix, $\mathbf{f} \in \mathbb{R}^k$ the latent factors with $\mathrm{Cov}(\mathbf{f}) = \Phi$, and $\boldsymbol{\varepsilon}$ benchmark-specific uniquenesses with diagonal $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \Psi$ and $\mathrm{Cov}(\mathbf{f}, \boldsymbol{\varepsilon}) = \mathbf{0}$. The decomposition of Σ into a common part $\Lambda\Phi\Lambda^\top$ and a unique part Ψ is the reason we use factor analysis in preference to principal components for a validity question. A principal component maximises total variance and so absorbs benchmark-specific noise along with shared capability, whereas the common-factor model attributes to the factors only the variance that benchmarks share (Thurstone, 1947). Parameters are estimated by maximum likelihood under multivariate normality. The dominance statistic for H1 is computed on the unrotated solution, where $\Phi = I$ and each factor's contribution to common variance is its sum of squared loadings, the m-th diagonal entry of $\Lambda^\top\Lambda$,

$$s_1 = \frac{\sum_{j=1}^{p} \lambda_{j1}^2}{\sum_{m=1}^{k}\sum_{j=1}^{p} \lambda_{jm}^2} = \frac{(\Lambda^\top\Lambda)_{11}}{\mathrm{tr}(\Lambda^\top\Lambda)}.$$

On the complete-case grid the three factors of the exploratory extraction carry sums of squared loadings 7.79, 2.34, and 0.32, so $s_1 = 7.79/10.45 = 74.5\%$; the statistic is defined on that extraction because under $k = 1$ it equals one by construction, and the first principal component's 79.4% of total variance is the count-free companion figure. This is a factor-level share, distinct from a single benchmark's communality $\sum_m \lambda_{jm}^2$, a diagonal entry of $\Lambda\Lambda^\top$.

Dimensionality by parallel analysis. Let $\ell_1 \geq \cdots \geq \ell_p$ be the eigenvalues of the sample correlation matrix R, and let $\ell_m^{(0.95)}$ be the 95th percentile of the m-th eigenvalue over correlation matrices of independent standard-normal data of the same dimensions $n \times p$. Parallel analysis retains factor m when $\ell_m > \ell_m^{(0.95)}$ (Horn, 1965). The rule asks whether a factor explains more than chance sampling structure would, so it is more conservative than the Kaiser rule $\ell_m > 1$ (Kaiser, 1960) whenever finite-sample noise pushes the random-data eigenvalues above one, as it does here ($\ell_2^{(0.95)} = 1.57$); on this battery both rules retain one factor. Information criteria such as the BIC (Schwarz, 1978) compare likelihoods of nested models and reward any factor that improves fit, which on a near-rank-one correlation matrix means retaining factors that carry very little common variance; the BIC's preference for $k = 4$ here (check C7, Appendix L) is an instance. On our grid $\ell_1 = 9.53$ against $\ell_1^{(0.95)} = 1.79$, and $\ell_2 = 0.91$ against $\ell_2^{(0.95)} = 1.57$, so one factor is retained (Figure 5).

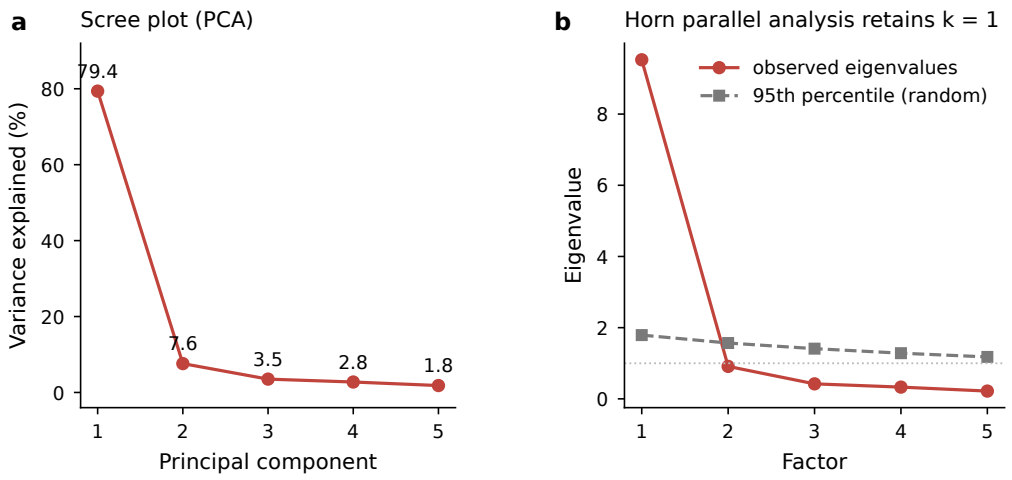


Figure 5: **Dimensionality of the battery.** (a) The scree plot of principal-component variance shares drops sharply after the first component (79.4%). (b) Horn parallel analysis retains a single factor, as only the first observed eigenvalue (9.53) exceeds its random-data 95th percentile (1.79), while the second (0.91) already falls below (1.57).

Oblique rotation. The maximum-likelihood solution is identified only up to rotation. Direct oblimin (Jennrich & Sampson, 1966) chooses the rotation that minimises

$$Q(\Lambda) = \sum_{m<l} \Big(\sum_{j=1}^{p} \lambda_{jm}^2 \lambda_{jl}^2 - \frac{\gamma}{p} \sum_{j=1}^{p} \lambda_{jm}^2 \sum_{j=1}^{p} \lambda_{jl}^2 \Big),$$

with $\gamma = 0$ giving the quartimin criterion we use, while allowing $\Phi \neq I$. An orthogonal rotation such as varimax forces $\Phi = I$; on data with a strong general factor this distributes shared variance across artificially independent axes, so that no single factor can carry the general variance and the dominance we set out to measure is obscured by construction. The date-adjusted factors correlate at 0.67 and 0.72 (Section 5.3), which confirms that the oblique choice was the right one.

Factor scores and the leakage-safe projection. Regression (Thurstone) factor scores are the linear predictor $\hat{\mathbf{f}} = W^\top \mathbf{z}$ that minimises $\mathbb{E}\|\mathbf{f} - W^\top \mathbf{z}\|^2$. Setting the derivative to zero gives $W = \Sigma^{-1}\mathrm{Cov}(\mathbf{z}, \mathbf{f}) = R^{-1}\Lambda\Phi$, where R is the sample correlation matrix, Λ the pattern loadings, and Φ the inter-factor correlation matrix of the sorted solution, so that $W = R^{-1}\Lambda$ when the factors are orthogonal (Thurstone, 1947). The structural factor scores behind H2 and the model clustering use the oblique form. The prediction task uses the orthogonal-form weights $W = R^{-1}\Lambda$ built from the pattern loadings; there W is estimated on the training fold only and applied to the held-out fold, the leakage-safe projection of Section 4.2. The two score sets are an invertible reparameterisation of one another, $\hat{\mathbf{f}}_{\text{oblique}} = \Phi\, \hat{\mathbf{f}}_{\text{pattern}}$, so an unpenalised regression on either gives identical fitted values; the ridge learners of Task 2 are not invariant to the reparameterisation, and the reported ladder uses the pattern form throughout. W is a function of the data, so if it were estimated on all n models, every held-out model's score would have contributed to the weights used to predict it, and the predictive comparison in Task 2 would be contaminated by exactly the information it is meant to withhold.

Date adjustment as removal of a rank-one component. Each standardised benchmark z_j is regressed on release date d (and, in a subsample, additionally on log training compute),

$$z_j = \beta_{0j} + \beta_{1j}\, d + r_j,$$

and the factor model is re-fitted on the re-standardised residuals r_j. The regression is linear in d, so the adjustment removes the linear component of the date trend; residualising on the logistic form that fits the date association best is a check we did not run. Writing $\mathbf{b} = (\beta_{11}, \ldots, \beta_{1p})^\top$, the covariance of the residual vector is $\Sigma_r = \Sigma - \sigma_d^2\, \mathbf{b}\mathbf{b}^\top$, so the adjustment removes from the benchmark covariance

exactly the rank-one component that lies along the date vector. If the general factor were nothing but the shared time trend, Λ_1 would be proportional to $\mathbf{b}$ and the first factor's share would collapse after adjustment; if it were unrelated to date, the share would be unchanged. The observed change, $74.5\% \to 59.6\%$, is the H2(ii) statistic and measures where the battery sits between those two cases. The design mirrors the scaling-law residualisation of Kearns (2026), with release date in place of parameter count as the variable adjusted for.

Nested leave-one-benchmark-out estimator. For an economic target t, the out-of-fold error of rung r is

$$\mathrm{MSE}_r = \frac{1}{n}\sum_{i=1}^{n}\left(z_{it} - \hat{z}_{it}^{(-\kappa(i))}\right)^2,$$

where $\hat{z}_{it}^{(-\kappa(i))}$ is the prediction for model i from a learner trained on the outer folds excluding the fold $\kappa(i)$ that contains i, with hyperparameters chosen by an inner five-fold grid search on those training folds alone. Selecting hyperparameters on the same folds that supply the reported error biases the estimate optimistically, and the bias grows with the size of the search space (Varma & Simon, 2006); nesting removes it at the cost of a five-fold increase in computation. The predictive-validity statistic contrasts the single-index rung against the k-factor rung, $\Delta\mathrm{MSE} = \mathrm{MSE}_{\mathrm{ii}} - \mathrm{MSE}_{\mathrm{iv}}$.

Paired bootstrap for ΔMSE. Let e_i^{ii} and e_i^{iv} be model i's out-of-fold squared errors under the two rungs. A 95% interval for $\Delta\mathrm{MSE}$ is obtained by resampling models with replacement, recomputing $\frac{1}{n}\sum_i(e_i^{\mathrm{ii}} - e_i^{\mathrm{iv}})$ on each of $B = 2000$ resamples, and taking the 2.5th and 97.5th percentiles (Efron & Tibshirani, 1993). Resampling the paired differences matters because the two rungs are evaluated on the same models and their errors are strongly correlated, whereas resampling the two error vectors separately would discard that pairing; the paired design removes the between-model variance that the two rungs share, so the interval reflects only the variance of the contrast. H4 is supported when this interval excludes zero. Because the resampling unit is a model, near-duplicate configurations of one base model would be treated as independent draws; the deduplicated re-run in Appendix L, whose resampling unit is a distinct base model, is the check on that assumption.

E DATE-ADJUSTED THREE-FACTOR LOADINGS

Supports Section 5.3. Table 3 tabulates the loadings shown as a heatmap in Figure 2b. F1 is an agentic and work-realistic factor, F2 an academic-knowledge factor, and F3 a physics and hard-reasoning factor; all four economic benchmarks load highest on F1.

Table 3: **Date-adjusted three-factor oblimin loadings**, economic benchmarks in bold.

Benchmark	F1	F2	F3
GDPval (Elo)	**0.84**	0.13	−0.04
Terminal-Bench v2.1	**1.01**	−0.12	0.03
τ^3-Banking	**0.54**	0.04	0.38
τ^2-Bench	**0.50**	0.31	−0.13
GPQA Diamond	0.01	0.95	0.04
HLE	0.32	0.28	0.49
AA-Omniscience	0.69	0.12	0.14
SciCode	0.26	0.61	0.20
CritPt	0.02	−0.00	0.99
Terminal-Bench Hard	0.82	0.10	0.10
AA-LCR	0.34	0.67	−0.12
IFBench	−0.13	0.86	0.02

F BENCHMARK BATTERY

Supports Section 3. Table 4 describes every candidate benchmark, the construct it targets, the type of score on which the leaderboard reports it, its coverage before and after the inclusion rule, and the role the rule assigned, with the source paper for each listed beneath. The economic benchmarks differ

from the academic ones in kind as well as in coverage. They score multi-step, tool-using, or human-preference-judged task completion on work-like tasks, whereas the academic benchmarks score single-response accuracy on knowledge or reasoning items. Three of the twelve (AA-Omniscience, AA-LCR, and Terminal-Bench Hard) are constructed or subset by the leaderboard operator itself; all twelve are run by that operator under a single harness (Artificial Analysis, 2026).

Table 4: **Candidate benchmarks by capability block**, with the construct each targets, the type of score reported, the number of models scored on the raw 548-configuration snapshot and among the 421 retained, and the role assigned by the inclusion rule. Twelve benchmarks are primary; MMMU-Pro is sensitivity-only for sparse and multimodal coverage; APEX-Agents falls below the sixty-model floor. Higher is better throughout; sources are listed below the table.

Benchmark	Construct measured	Score type	Raw	Retained	Role
Economic					
GDPval (Elo)	Paid professional task quality	pairwise-judged Elo	117	112	primary
Terminal-Bench v2.1	Terminal and agentic task completion	pass rate	121	121	primary
τ^3-Banking	Multi-step banking workflows	task success	112	112	primary
τ^2-Bench	Tool-use agentic tasks	task success	428	413	primary
APEX-Agents	Agentic task completion	task success	26	–	dropped (sparse)
Academic					
GPQA Diamond	Graduate-level science QA	accuracy	513	421	primary
HLE	Frontier academic knowledge	accuracy	509	421	primary
AA-Omniscience	Broad knowledge, hallucination-penalised	penalised score	418	418	primary
MMMU-Pro	Multimodal understanding	accuracy	204	–	sensitivity only
Scientific coding					
SciCode	Research-level scientific coding	pass rate	507	421	primary
CritPt	Frontier physics problem solving	accuracy	422	419	primary
Terminal-Bench Hard	Hard terminal coding tasks	pass rate	421	413	primary
Long-context / instruction					
AA-LCR	Long-context reasoning	accuracy	441	420	primary
IFBench	Instruction following	accuracy	437	417	primary

Sources. GDPval (Patwardhan et al., 2026); Terminal-Bench v2.1 (Merrill et al., 2026); τ^3-Banking (Shi et al., 2026); τ^2-Bench (Barres et al., 2026); GPQA Diamond (Rein et al., 2024); HLE (Phan et al., 2025); SciCode (Tian et al., 2024); CritPt (Zhu et al., 2025); IFBench (Pyatkin et al., 2025); AA-Omniscience, AA-LCR, and Terminal-Bench Hard are constructed or subset by the leaderboard operator (Artificial Analysis, 2026).

GDPval's Elo rating comes from pairwise, head-to-head judgements of model outputs on economically valuable tasks, so a higher rating means more often preferred on a scale with no fixed maximum. The τ-Bench family (Yao et al., 2025; Barres et al., 2026) scores an agent on whether the final database state and required confirmations match a policy-compliant goal after a multi-turn conversation with a simulated user; τ^3-Banking is the banking-knowledge domain of its third generation, in which agents must complete policy-compliant state changes over a large document base. The leaderboard labels its Terminal-Bench column v2.1; the cited paper presents the suite's 2.0 release. MMMU-Pro (Yue et al., 2025) is demoted to a sensitivity-only role and APEX-Agents (Vidgen et al., 2026) is dropped for sparsity (Appendix G).

G Full coverage audit

Supports the inclusion rule of Section 3. Table 4 records every candidate benchmark's coverage on the raw snapshot and the decision applied to it, and Figure 6 shows that coverage together with the pairwise overlaps that set the effective sample size for every correlation in the study. The sixty-model floor is the smallest count at which the sparsest retained benchmark still overlaps every other benchmark on more than one hundred models (Figure 6b), which keeps each pairwise correlation estimable with a standard error below about 0.1.

H Exploratory data analysis

Supports Section 3. We examined coverage and missingness, correlation structure, and temporal patterns before fitting any model (Figure 7). All standardisation and model fitting uses the training

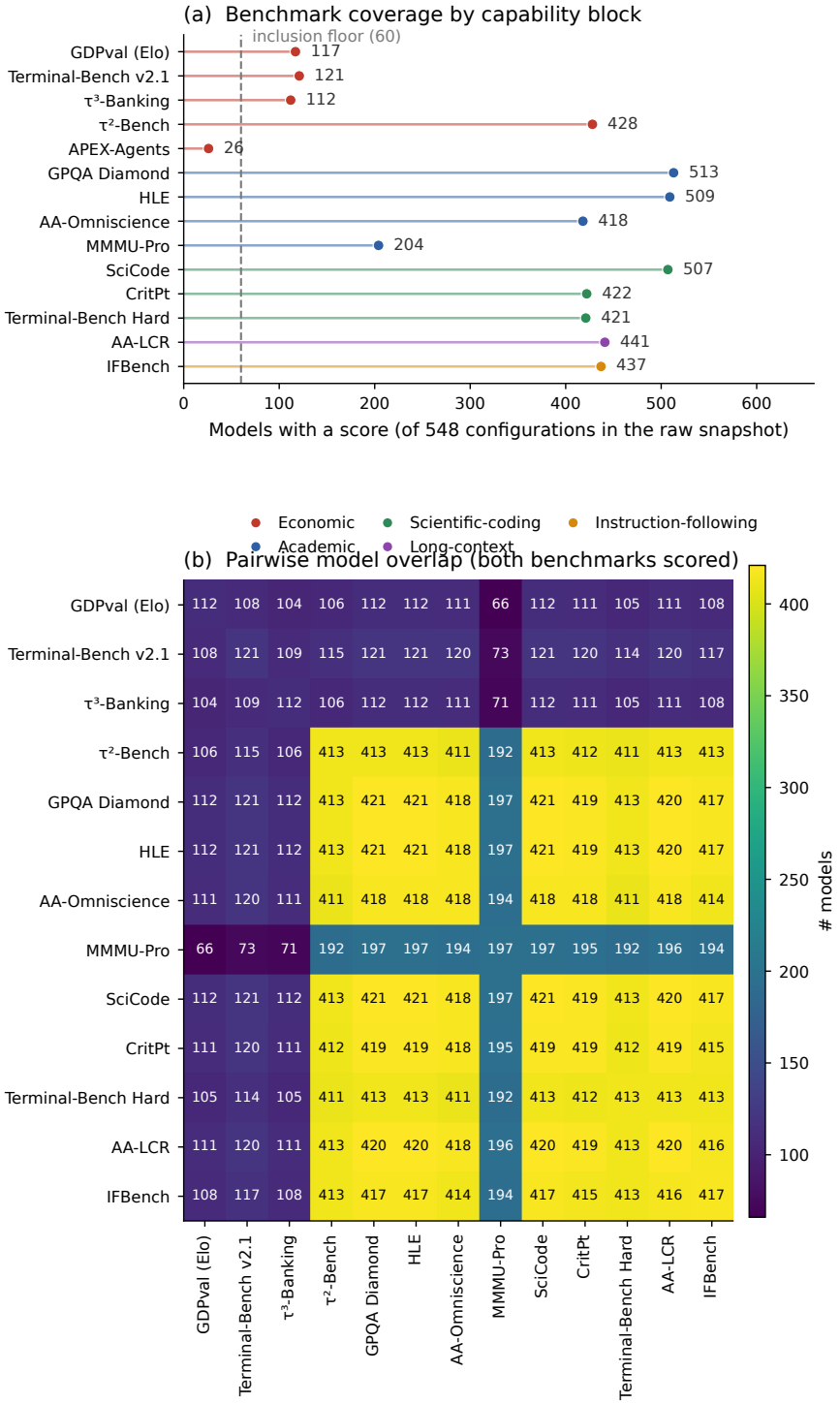


Figure 6: **Coverage audit behind the inclusion rule.** (a) Number of models carrying a score on each candidate benchmark on the raw 548-configuration snapshot, coloured by capability block, with the sixty-model inclusion floor marked. Three economic benchmarks (Terminal-Bench v2.1, GDPval, and τ^3-Banking) are scored on only 112 to 121 models, while τ^2-Bench and the academic core are each scored on four hundred or more; APEX-Agents, with 26 models, falls below the floor. (b) Pairwise counts of models scored on both benchmarks in a pair, which set the effective sample size for every correlation in the study and are the evidence for the two-grid design.

portion of each split only, so these views characterise the data without leaking test information into the models.

The battery is highly collinear before any modelling. The mean off-diagonal Spearman correlation is $\rho = 0.79$, every benchmark pair is positive, and the economic benchmarks correlate with the academic core almost as strongly as with one another (Figure 7a). A dominant general factor would produce this near-block structure, and the structure analysis must explain it, since a strong positive manifold is equally consistent with one underlying capability and with several capabilities that improve together over time. Scores rise steeply and jointly with release date (Figure 7b), so release date bears on any claim that two benchmarks measure the same thing; the saturation literature finds saturation rising with benchmark age (Akhtar et al., 2026), and we therefore adjust for date before reading the factor structure.

Most benchmarks are bounded accuracies in $[0, 1]$, and several are strongly right-skewed, with CritPt at Fisher skewness 3.8, HLE at 1.5, and τ^3-Banking at 0.9, their scores piling up near the floor (Figure 8). Two benchmarks are unbounded, GDPval on an Elo scale and AA-Omniscience on a penalised score, and both sit closer to symmetric. Bounded and skewed scales can manufacture or hide structure through the metric alone (Schaeffer et al., 2023), which motivates the logit-transform check in Appendix L; that check confirms that the first-factor share is no artefact of the raw scale. The low-scoring tails hold weak models whose scores are low, so we retain every row and handle the boundedness through the transform, trimming no outliers.

The dimensionality evidence behind H1, the scree plot and the parallel analysis, is Figure 5 in Appendix D; the first principal component accounts for 79.4% of total score variance, and only the first observed eigenvalue exceeds its random-data 95th percentile.

I Hyperparameter grid and selections

Supports Section 4.2 and the robustness claim of Section 5.5 that the H4 verdict does not hinge on a single tuning choice. The inner grid searched ridge $\alpha \in \{0.03, 0.1, 0.3, 1, 3, 10, 30\}$, elastic-net $\alpha \in \{0.03, 0.1, 0.3, 1\}$ with ℓ_1 ratio in $\{0.2, 0.5, 0.8\}$, random-forest depth in $\{4, \text{none}\}$ with 300 trees, and gradient-boosting depth in $\{2, 3\}$ with learning rate in $\{0.05, 0.1\}$. The ridge and elastic-net grids span three orders of magnitude of regularisation strength on a log scale, the standard practice for a regularisation path (Hastie et al., 2009), and the tree grids bracket the shallow trees appropriate for $n = 96$ and $k + 4$ predictors. Table 5 reports the modal selection per rung; the modal ridge $\alpha = 0.03$ sits at the weakly regularised end of the grid, and the H4 verdict is unchanged across the sweep of k in Appendix L. The regression learners and cross-validation use scikit-learn (Pedregosa et al., 2011); the factor analysis uses the `factor_analyzer` package, and SHAP its reference implementation (Lundberg & Lee, 2017).

Table 5: Modal hyperparameters selected by the inner five-fold grid search, reported as the plurality choice across all twelve leave-one-benchmark-out targets with the agreement count. The H4 verdict rests on fold-averaged out-of-fold errors and on no single setting.

Rung	Learner	Modal hyperparameters	Agreement
(i) timing / scale	random forest	`max_depth=4, min_samples_leaf=3, n_estimators=300`	8/12
(ii) mean index	ridge	`alpha=0.03`	6/12
(iii) first factor	elastic net	`alpha=0.03, l1_ratio=0.2`	5/12
(iv) k factors	ridge	`alpha=0.03`	6/12
(v) k factors + covariates	ridge	`alpha=1`	6/12

J Deviations from the pre-specified design

Supports the reproducibility claims throughout. A study that fixes its analysis plan in advance owes the reader an explicit account of where the delivery departed from it. Six departures matter. First, the plan's primary unit was one row per base model at its best default configuration. The analysis instead uses all 421 configurations; the deduplicated base-model dataset appears in Appendix L as planned check R1, and the configuration level serves as the primary. Second, the plan's factor-count rule is parallel analysis, and it retains one factor. The three-factor solution that exhibits economic separation

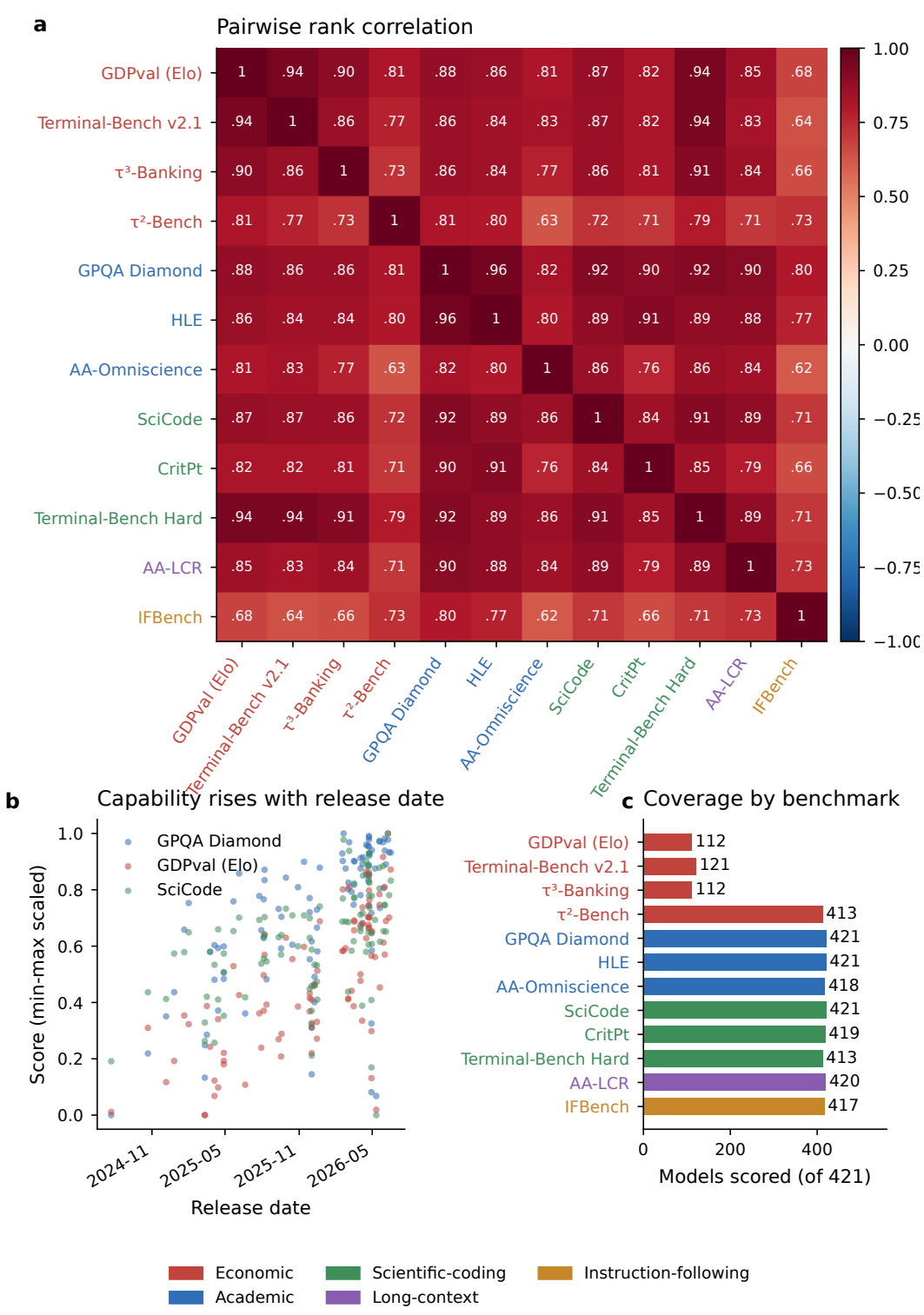
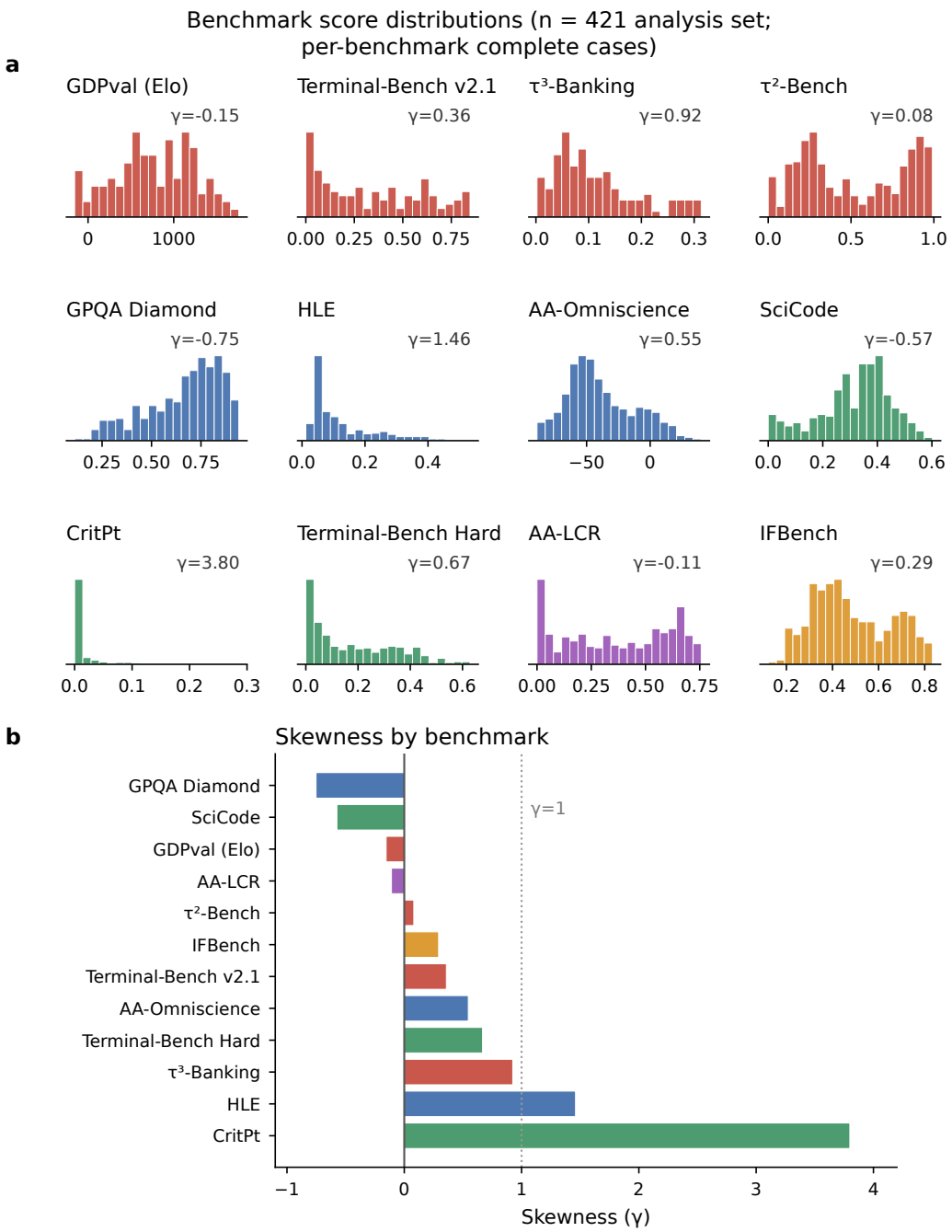


Figure 7: **Exploratory data analysis of the twelve-benchmark battery on the complete-case grid ($n = 96$).** (a) Pairwise Spearman rank correlations, with benchmark labels coloured by capability block; almost every pair is strongly positive. (b) Per-benchmark scores (min–max scaled) against model release date, showing capability rising steeply with time. (c) Coverage per benchmark across the 421 retained models, with the four economic benchmarks (red) scored on far fewer models than the academic core.



(a) Raw score histograms by taxonomy block; γ = Fisher skewness. (b) CritPt, HLE, and τ³-Banking are strongly right-skewed (γ > 0.9), motivating the logit robustness check on bounded accuracies.

Figure 8: **Univariate benchmark score distributions.** (a) Raw score histograms for the twelve benchmarks, coloured by capability block; γ is the Fisher skewness. (b) Skewness ordered across benchmarks. Most benchmarks are bounded accuracies in $[0, 1]$, and several are strongly right-skewed (CritPt $\gamma = 3.8$, HLE $\gamma = 1.5$, τ^3-Banking $\gamma = 0.9$), with scores concentrated near the floor. GDPval (Elo) and AA-Omniscience are on unbounded scales and are closer to symmetric.

is a pre-declared exploratory over-extraction, and we report H3 as not supported under the plan in consequence. Third, the taxonomy places Terminal-Bench Hard in scientific coding, while the plan grouped it with Terminal-Bench v2.1 as economic. The date-adjusted loadings show Hard at 0.82 on the economic factor, so the assignment is contestable, and we disclose the split openly. Fourth, the delivered robustness checks depart from the planned suite. We label them distinctly below, and we do not renumber a different battery onto the planned names. Fifth, the plan set the two inclusion floors as indicative values for confirmation at the coverage audit; the audit confirmed both unchanged, and no score was used to choose them. The plan also targeted 180 to 220 base models, where coverage across all twelve benchmarks delivered 96 complete cases and 89 deduplicated base models. Sixth, the plan specifies $B = 1{,}000$ bootstrap resamples, names both the timing baseline and the single index for H4, and fixes no learner for the contrast; we use $B = 2{,}000$, test against the single index as the stricter baseline, and report ridge with the other three learners in check C8.

Two further differences are presentational and are recorded for readers comparing the paper with the deposited plan. The plan organised the work under two questions, structure and prediction; the paper splits the structural question so that each hypothesis has its own research question, with no change to any hypothesis or test. And the H2(ii) fifteen-point threshold was calibrated to a reading of Kearns (2026) in which a scale correction lowered a dominant factor's share from 72% to 41%. That reading was mistaken, and Section 2 gives the corrected account; the threshold itself was fixed in advance and has not been moved.

K Clustering of models and benchmarks

Supports Section 4.1. A factor model can impose structure the raw data do not support, so we sought a check that does not depend on the retention rule, and clustering in factor-score space corroborates the low dimensionality. We cluster models in factor-score space with k-means, Gaussian mixtures, and agglomerative Ward linkage, selecting the cluster count by silhouette (Rousseeuw, 1987) and the gap statistic (Tibshirani et al., 2001); three algorithms guard against a clustering that reflects one algorithm's inductive bias. All three prefer $k = 2$ by silhouette, at 0.480 for k-means, 0.499 for agglomerative, and 0.467 for the Gaussian mixture (Figure 9b). The gap statistic increases monotonically in k and finds no interior optimum, so it is inconclusive on data with a strong general factor and we set it aside. The two clusters mark capability tiers along one axis of overall strength. A larger, earlier, lower-scoring group of 302 models, with median release 2025-09 and mean Intelligence Index 13.7, separates from a smaller, newer, higher-scoring group of 107 models, with median release 2026-03 and mean index 37.1. The split sorts models by how much capability they hold, along the same date-driven axis (Figure 9a).

Clustering the benchmarks separately by correlation distance $1 - \rho$ under average linkage places ten of twelve, and three of four economic benchmarks, in one block, with τ^2-Bench and IFBench branching separately (Figure 10).

L Robustness checks

Supports Section 5.5. We first report the one planned check that bears directly on the deduplication concern, then the additional checks actually delivered.

Planned check R1 (configuration versus deduplicated). We collapse the 421 configurations to one row per base model, keeping the best default configuration as the plan specified. We operationalise it as the highest-intelligence-index row for each base model, after stripping reasoning, effort, and preview suffixes from the model name. Of the resulting base models, 89 carry the complete battery. On this deduplicated grid the first-factor share rises to 90.2% and the date-adjustment drop rises to 24.1 points, and the four economic loadings remain above 0.40. Deduplication therefore strengthens H1 and H2(ii) and leaves the exploratory economic pattern intact, and the configuration-level analysis is the conservative choice for the structural headline. We also re-run the H4 predictive test on this deduplicated grid, where the resampling unit is a distinct base model. The pooled economic-block improvement of the k-factor representation over the single index is $+0.038$ with a 95% interval of $[+0.020, +0.056]$. It matches the configuration-level $+0.037$ and again excludes zero, so the predictive gain does not depend on treating near-duplicate configurations as exchangeable.

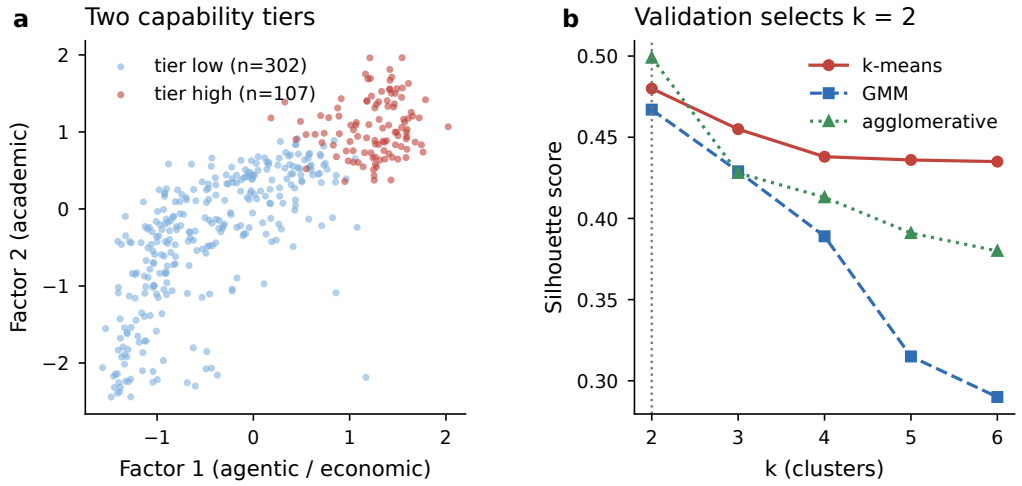


Figure 9: **Model clustering on the dense grid ($n = 409$).** (a) Models in the space of the first two factors, coloured by the two-cluster agglomerative solution. The factor axes are the dense-grid solution fitted largely without the economic columns, so they track overall capability, and the two clusters mark tiers of it. (b) Silhouette scores by k; all three algorithms agree on $k = 2$.

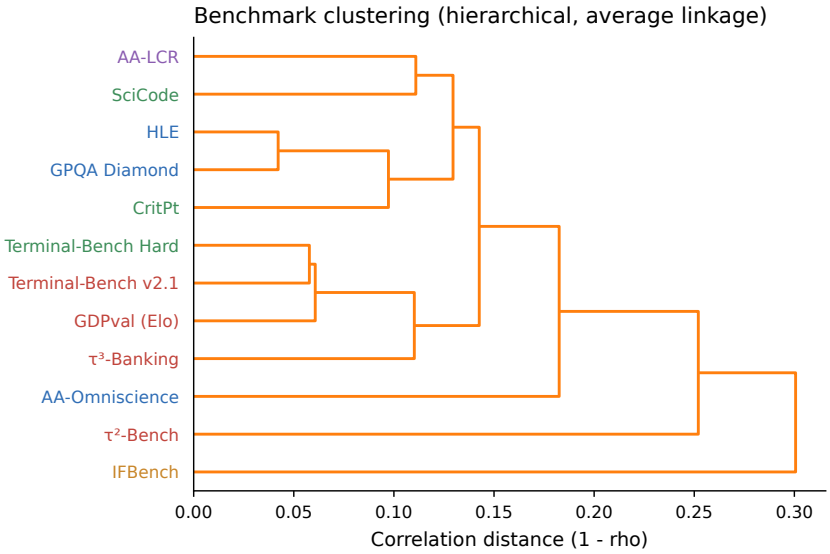


Figure 10: **Benchmark clustering by correlation distance** ($1 - \rho$, average linkage). Ten of the twelve benchmarks, including three of four economic benchmarks, fall in one block, with τ^2-Bench and IFBench branching separately.

Delivered checks C1–C8. (C1) A rank-based factor analysis raises the first-factor share to 94.5%, and (C2) a logit transform of the bounded accuracy benchmarks raises it to 91.8%, so H1 is not a Pearson artefact. (C3) On the 58-model compute-known subsample, the date-only adjustment drops the first-factor share by 16.5 points, and adding log-compute jointly gives a smaller 9.3-point drop, so compute does not deepen the temporal correction beyond date on the same models. (C4) The three clustering algorithms agree on $k = 2$. (C5) The predictor-ladder ordering holds on both the economic block and all twelve targets. (C6) The dominant-factor date fit does not depend on the link function, since a logistic fit gives $R^2 = 0.505$ against an ordinary-least-squares $R^2 = 0.477$, and both clear 0.30 with the first factor ranked strongest, so H2(i) does not depend on that modelling choice. (C7) The k-factor predictive advantage does not depend on the exploratory factor count. Sweeping the

rung-(iv) predictor over $k \in \{2, 3, 4, 5\}$, the improvement over the single index is positive at every count with a bootstrap interval excluding zero, reading $+0.021$ at $k = 2$, $+0.037$ at the $k = 3$ we use, $+0.058$ at the criterion's $k = 4$, and $+0.059$ at $k = 5$, with pooled test R^2 rising from 0.79 to 0.83 across the range. The gain increases with k, so the $k = 3$ we adopt is conservative for this test relative to the criterion's $k = 4$. (C8) The k-factor advantage depends on the learner family. Fixing each learner on both rungs, the pooled economic ΔMSE is $+0.037$ for ridge and $+0.036$ for elastic net, positive on all four targets for both, and -0.23 for random forest and -0.18 for gradient boosting, negative on all four. The two tree learners fit the economic block worse on both rungs (test RMSE 0.49 and 0.53 on the index rung against ridge's 0.47) and degrade further with three factor inputs on 96 models, so the advantage holds for the linear learners that fit best and reverses for the tree learners; intervals were computed for ridge only.

The planned checks not run are R3 (FIML and MICE missingness sensitivity), R4 (replication of the single-factor foil then correction), R5 (reasoning-effort variants as a second scale axis), and R6 (the taxonomy swap test). Each was deferred for space; R5 and R6 bear on the effort and taxonomy departures above and come next.

M Explainability and error analysis

Supports Section 5.4. To interpret the predictive gain, we open the models fitted to the largest economic target. GDPval is predicted best by a k-factor ridge with out-of-fold R^2 0.88 (Figure 11). SHAP attribution on the gradient-boosting learner fitted at the same rung ranks the agentic factor F1 first, followed by the reasoning flag and the hard-reasoning factor F3, with scale and open-weights status contributing little (Figure 3c). The ordering is informative in the light of prior work, since Ruan et al. (2024) find agentic performance forecastable from a low-dimensional capability space, and Ilić & Gignac (2024) find parameter count correlated with the general factor at about 0.6; here, once the factors are in the model, parameter count and open-weights status add almost nothing, so the scale signal that prior work reports reaches GDPval through the factors alone. The predicted-versus-observed plot lies on the identity line, and the residuals show no systematic trend (Figure 11a,b). The largest residuals are interpretable outliers. They include two non-reasoning variants that outperform their economic predictions, Grok 4.3 Non-reasoning at $+1.14$ and Qwen3.6 27B Non-reasoning, and six models that under-perform theirs, led by Gemini 3.1 Pro Preview and Llama 4 Maverick (Figure 11c). The pattern fits GDPval rewarding agentic behaviour that the other benchmarks capture only indirectly.

N Data provenance and reproducibility

Supports Section 3 and the reproducibility claims throughout. The primary data set is the Artificial Analysis model leaderboard (Artificial Analysis, 2026), captured 2026-07-06 as a raw HTML snapshot and pinned by the SHA-256 hash recorded in the shipped manifest (`aa_snapshot_manifest.json`), `6f19f8f08befaaf14cb2d9952e379404c39f246b52d7ae4d1c2aa44795404de6`. The Data API refused an unauthenticated request, so we captured the public page and stored the exact bytes. Every downstream number is therefore reproducible regardless of live-page drift (Biderman et al., 2024). The secondary data set is Epoch AI Notable AI Models (Epoch AI, 2026), released under CC-BY 4.0 and used only for the scale robustness check. The analysis plan was fixed before the data were captured and before any analysis was run, and is deposited in a public registry, which does not support anonymised links for completed deposits; the deposit is retrospective, and the note there records the timeline, the plan's checksum, and the limits of that evidence. The recorded timeline is that the plan was first saved at 14:18 UTC on 6 July 2026, the leaderboard snapshot was captured at 15:17 and the first analysis artefact was produced at 15:23, so the plan precedes the data by 59 minutes and the analysis by 65; the plan's three recorded copies are byte-identical, and the timestamps come from the working environment's own file record, which is weaker evidence than a registry timestamp would be. Both snapshots, every result table, and the analysis notebook are released in an anonymised repository (link in the supplementary material). Reproducibility is tiered, and each result belongs to one tier. The structural results, the SHAP attribution, the deduplication checks, and the prediction ladder are *recomputed* from the pinned inputs; twelve result tables and four appendix figures are *regenerated* from archived outputs, which the notebook reads and prints without recomputing. The pipeline is

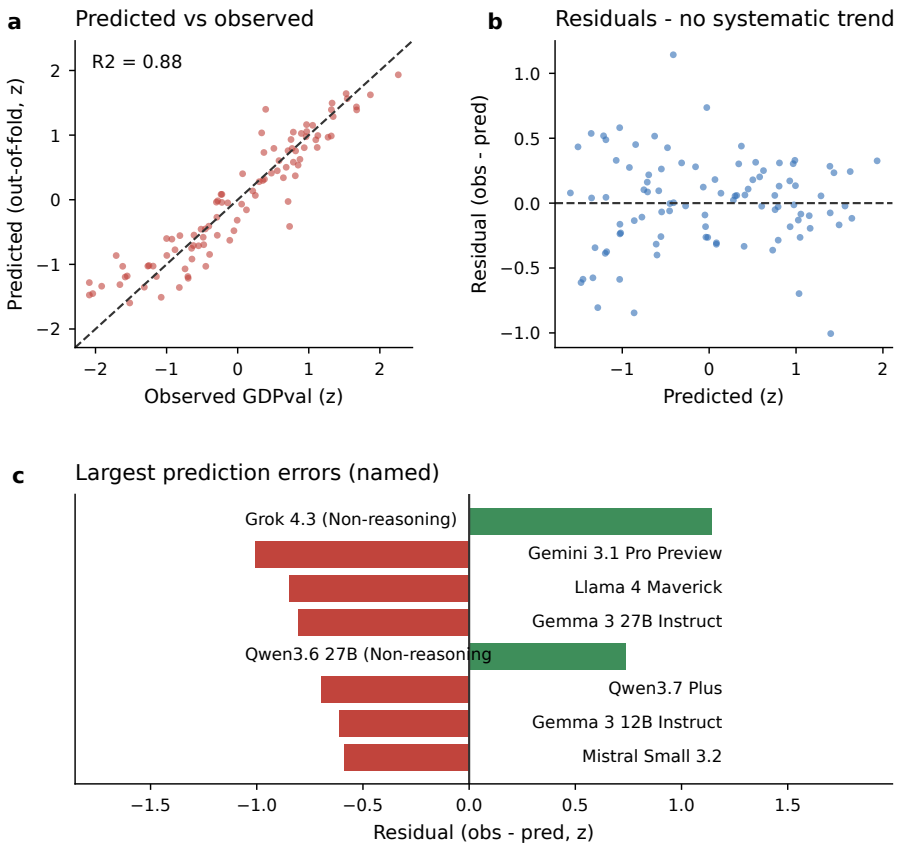


Figure 11: **Regression diagnostics for the GDPval k-factor ridge model, out-of-fold.** (a) Predicted versus observed, on the identity line with $R^2 = 0.88$. (b) Residuals versus predicted show no systematic trend. (c) The eight largest residuals, named. These are interpretable outliers, and none signals misspecification.

a single notebook that runs top to bottom from the pinned snapshots with no manual step, no live network call and no path outside the repository, and all four learners run by default. It regenerates the figures that carry the results, Figures 2, 3, 7, 9 and 11, while the remaining figures and twelve result tables listed in the repository are shipped without being recomputed within it; the notebook prints every value it reads, so each remains checkable against its own output. The SHAP attribution behind Figure 3c and the R1 deduplication with its H4 re-run verify themselves against the tables cited here and reproduce exactly.

The prediction ladder is refitted on every run and written alongside the canonical table, so no value reported here can be altered by re-execution. The notebook then recomputes the quantities this paper prints, the per-target and pooled economic ΔMSE and the five economic-block rows of Table 1, and compares them with the canonical values at a tolerance of 5×10^{-4}, half a unit at the third decimal and so at the precision reported here; it separately verifies that the best-learner ranking behind the table's learner column is unchanged. Those quantities reproduce in every printed digit on both machines we have tested. Individual ladder cells are reported without being enforced, because one rung is not platform-invariant. The maximum-likelihood solution has a weakly determined leading direction for this correlation structure, so the single-factor rung can settle on a slightly different point under a different linear-algebra backend, and gradient-boosted trees carry that difference most visibly because a split threshold either flips or does not. The largest deviation we observed was 7.2×10^{-4}, reaching the k-factor rung only through tree learners on a non-economic target, while the mean-index and k-factor ridge fits from which every reported effect and interval is built agreed to about 10^{-7}. The divergence is a property of the platform and of that rung and leaves the results untouched. The released library (Anonymous, 2026), extracted from this notebook, recomputes the

structure results and the ladder point estimates from the same pinned snapshot by one command, statistically reproduces the bootstrap interval and the k-sweep under a fresh random stream, and regenerates the clustering table and three further fields from the archive; its ledger records the archived value, the recomputed value and the tolerance for every quantity, and 23 of its 44 acceptance tests lock this paper's numbers. Two of its computations reconstruct steps the archive does not ship, the four-parameter logistic behind the date R^2 and the mean-of-RMSE pooling behind the ladder table, and both are recorded as reconstructions in its decision record.