

How we test an AI that forecasts

What "right" means, how to measure it independently, and why the headline accuracy number is the least useful thing you will learn.

Most organisations now run at least one system that predicts something: next quarter's demand, the price a customer will accept, the probability a claim is fraudulent, the cash a project will need. More and more of these are AI tools bought from a vendor, whose workings the buyer cannot see and whose accuracy nobody independent has checked. The question we are asked is simple: is it right?

The answer is never a single number. "Right" means wrong by an acceptable amount, in the places where being wrong costs the most, consistently over time, and by less than a cheaper alternative would be. That definition is not new. It is the standard that forecasting science has applied to statistical models for decades^{1,2}, that bank supervisors apply to the models behind lending and capital decisions^{3,4}, and that the UK government applies to its own analysis⁵. AI tools have mostly escaped it, not because the standard does not fit, but because nobody has been asked to apply it. This note sets out how we do.

The seven steps



1

Start with the decision, not the model

Before we look at a line of output we ask what the forecast is used for and what a miss costs. A demand forecast feeding a reorder rule is punished for under-forecasting far more than over-forecasting. A credit score is judged on the cases it gets badly wrong, not its average. A pricing engine is judged on margin left on the table and business lost.

This determines the error measure. Mean absolute error, root mean squared error, scaled error against a naive forecast², directional accuracy, calibration of probabilities, cost-weighted loss: each answers a different question, and the wrong one can make a bad tool look good. Proper scoring rules exist precisely so that a model cannot game the measure it is judged on⁶. We write the chosen measures down, with the reasoning, before testing begins.

2

Build a reference set the model has never seen

A forecast can only be judged against outcomes that were unknown when it was made. That rules out most of what vendors show buyers, which is performance on data the model was built on or tuned against.

We assemble a reference set of real decisions and real outcomes, using the data exactly as it stood at the time of each forecast. Economists call this vintage, or real-time, data; it matters because figures get revised, and a model tested on revised history is being handed information it did not have⁷. We check for leakage, where a later fact has crept into an earlier input, and we hold the set back from the vendor and the model owner until the test is complete.

3

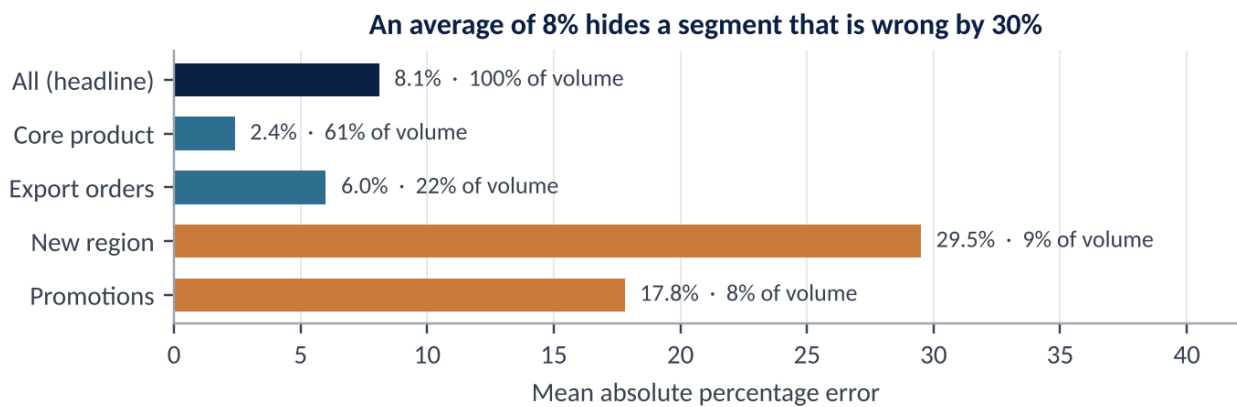
Set the benchmark

Every AI forecast is compared with something simple. For a time series that may be last period's value, a seasonal average, or a plain regression on two or three obvious drivers. For a classifier it may be a base rate or a scorecard an analyst could run in a spreadsheet. The benchmark is not a straw man. It represents what the organisation could have for almost nothing. If the AI cannot beat it by a margin that pays for its cost and complexity, that is the finding, however sophisticated the tool. Whether the margin is real rather than luck is itself a statistical question, with standard tests to answer it¹.

4

Measure the errors, then take them apart

Headline accuracy is the least useful number we produce. The useful work is decomposition: where is the model wrong, and why? We split errors by segment (product, region, customer type), by horizon (one week out versus one quarter), and by regime (stable periods versus shocks). We separate bias, a consistent lean in one direction, from noise.



Illustrative data. Amber = segments exceeding the tolerance set at scoping.

Figure 1. Error decomposition by segment. The headline figure is dominated by the core product, where the model performs well. The two segments added after the model was built carry most of the risk.

We also test whether the model's own uncertainty bands are honest. If it claims a 90% range, do outcomes fall inside it nine times in ten? Figure 2 shows the test. Over-confident models are the most dangerous kind, because they are trusted most exactly when they should be trusted least.

This is where most findings come from. A tool that is 8% off on average is often 2% off on the core business and 30% off on a segment nobody checked, and that segment is usually the one that was added after the model was built.

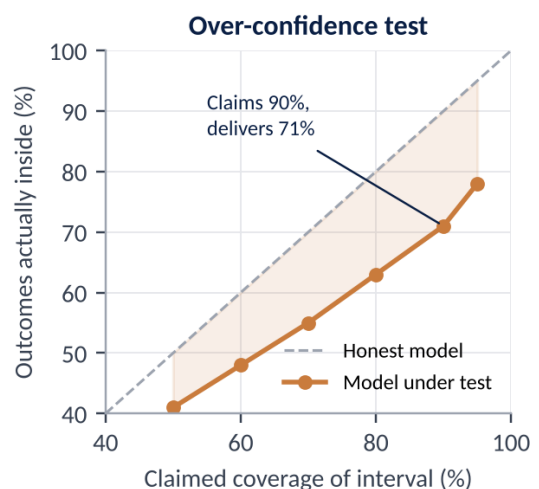


Figure 2. Calibration. Points below the diagonal mean the model's stated uncertainty is too narrow.

5

Test whether it holds up

A model that passed last year's test is not thereby reliable this year. We re-run the tests on rolling windows to see whether accuracy is drifting, and we look for structural breaks: points where the relationship between inputs and outcomes changed and the model did not notice. We stress the model with the conditions it will eventually meet, such as a demand collapse, a price spike or a shift in the customer mix, and report how it behaves when the world leaves the range it was trained on.

Where we can see the model's inputs we also check specification: whether the drivers make economic sense, whether any are proxies for the outcome itself, and whether the model is leaning on a variable that will not be available in time when the forecast is needed.

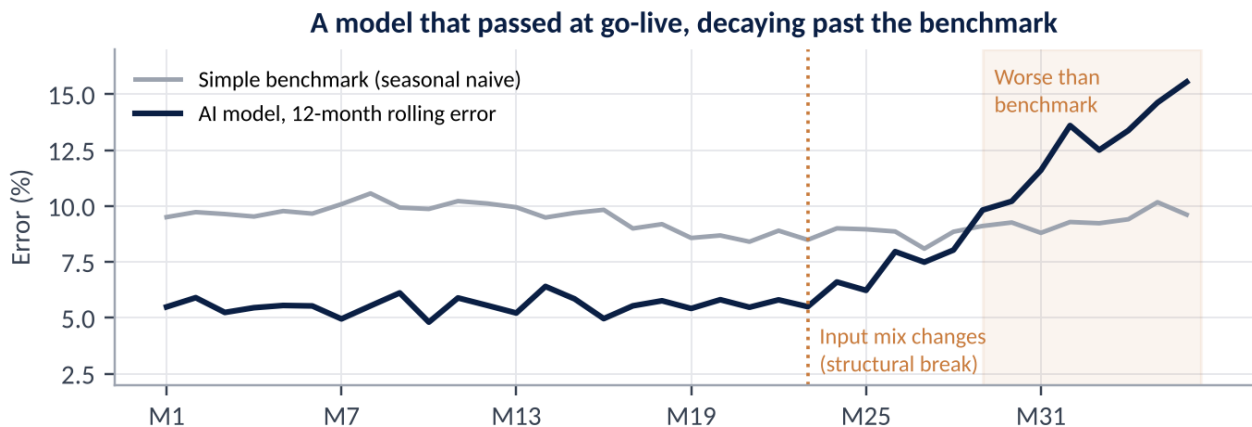


Figure 3. Rolling-window error. The model beat the benchmark comfortably at go-live. A change in input mix at month 22 began a decay that nobody was monitoring; by month 28 the simple benchmark was better.

6

Report what matters, rank it, sign it

The output is a short report, not a data dump. Findings are ranked by severity and by how much they bear on the decisions the model supports. Each carries the evidence, the likely cause and a recommendation: fix, constrain, monitor or replace. We state what the model is fit for and what it is not.

Every report ends with a signed opinion from the named reviewer. It says, in a paragraph a board or a regulator can read, whether the model is fit for its stated use and on what conditions. We put our name to it because an opinion nobody will sign is not an opinion. Independence matters here for the reasons the audit literature has documented at length: an assessor paid to find nothing usually finds nothing^{8,9}.

7

Come back

Models decay. We set a re-test schedule, typically quarterly for anything feeding a live decision, with a short monitoring pack that re-runs the core tests on new outcomes. Drift gets caught before it reaches the accounts.

What you receive

Engagement	Question answered	Deliverables	Duration
Output accuracy review	How often is the AI wrong, and where?	Error rates with confidence ranges; decomposition; ranked fixes; signed opinion	2–3 weeks
Model validation	Is the model well specified, stable, and better than a benchmark?	Backtests; benchmark comparison; stability and stress tests; signed report	2–4 weeks
AI impact evaluation	Did the AI deliver what it was bought for?	Effect separated from other factors, with ranges; evidence for the next decision	3–5 weeks
Monitoring	Is it still right?	Quarterly re-run of core tests on new outcomes; drift alerts	Ongoing

The constraint is almost always the reference set. If outcome data is thin or poorly recorded we say so early, because that is itself a finding about how the organisation is managing its AI. Where we report against frameworks we use ISO/IEC 42001¹⁰, the NIST AI Risk Management Framework¹¹ and DSIT's AI Management Essentials¹², and we quality-assure our own work to the principles of HM Treasury's Aqua Book⁵.

If you rely on an AI to forecast, price or decide and nobody independent has tested it against what actually happened, that is the place to start.

References and further reading

1. Diebold, F.X. and Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. The standard test of whether one forecast is better than another by more than chance.
2. Hyndman, R.J. and Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688. Why the choice of error measure matters, and the case for scaling against a naive forecast.
3. Board of Governors of the Federal Reserve System and OCC (2011). *Supervisory Guidance on Model Risk Management*, SR 11-7. The reference regime for independent model validation in banking.
4. Bank of England, Prudential Regulation Authority (2023). *Model risk management principles for banks*, Supervisory Statement SS1/23. The UK equivalent, explicitly covering AI and machine-learning models.
5. HM Treasury (2015). *The Aqua Book: guidance on producing quality analysis for government*. The UK standard for assured analysis: documented assumptions, version control, sensitivity testing, independent review.
6. Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. The theory behind judging probabilistic forecasts so that honesty is the best strategy.
7. Croushore, D. (2011). Frontiers of real-time data analysis. *Journal of Economic Literature*, 49(1), 72–100. Why testing on revised data overstates accuracy, and how to use vintage data instead.
8. Raji, I.D., Xu, P., Honigsberg, C. and Ho, D. (2022). Outsider oversight: designing a third party audit ecosystem for AI governance. *Proceedings of AIES '22*. arXiv:2206.04737. What financial and environmental audit teach about independence, access and conflicts of interest in AI assessment.
9. Partnership on AI (2026). *Demand and Incentives for External AI Assurance*. Why organisations under-buy independent assurance, and what changes that.
10. ISO/IEC 42001:2023. *Information technology — Artificial intelligence — Management system*.
11. National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1.
12. Department for Science, Innovation and Technology (2025). *AI Management Essentials*; and (2024) *Assuring a Responsible Future for AI*, with economic analysis by Frontier Economics. The UK government's self-assessment tool and its survey of the UK AI assurance market.
13. Centre for the Governance of AI and co-authors (2026). *Frontier AI Auditing: Toward Rigorous Third-Party Assessment*. arXiv: 2601.11699. Proposes graded AI Assurance Levels; useful for calibrating how much evidence a given claim needs.

Grant Hixon is an econometrician with twenty years building and testing forecasting and risk models at Dun & Bradstreet, the EIU, Continuum Economics, Oxford Economics and CRU. BlackPoint Economics provides independent AI and model assurance, economic evaluation and commissioned analysis. [**info@blkpoint.co.uk**](mailto:info@blkpoint.co.uk)

Figures 1–3 use illustrative data constructed to show the pattern of each test; they are not client results. © 2026 BlackPoint Ltd. Registered in England & Wales, company no. 16981186.