HOLDOUT VALIDATION STUDY

Retention Stability Engine

Independent Temporal Holdout · N = 51 Candidates · April 2026

84.3%	72.5%	0.169
12-Month Accuracy	Score Within ±15 pts	Mean Brier Score

This report presents results from an independent temporal holdout validation of the Ros Retention Stability Engine. Candidate outcomes were withheld during scoring. Results reflect how the engine performed against outcomes it had not seen.

Cohort size: 51 candidates Study period: Q1 2026 Freeze date: April 8, 2026
Study type: N-1 temporal holdout (prior history only — target role withheld)

Executive Summary

The Ros Retention Stability Engine was subjected to an independent temporal holdout study across 51 candidates. Each candidate's most recent role was withheld from the model during scoring — the engine saw only prior job history and was asked to predict retention risk for the hidden role. Actual employment duration in the withheld role served as ground truth.

The study was designed to simulate real-world deployment: a hiring team presenting a candidate file before an offer is extended. The model receives no information about the outcome it is predicting. This eliminates look-ahead bias and provides a clean pre-hire signal test.

Metric	Result	Context
Cohort size	**51 candidates**	Best-per-candidate hybrid holdout
Score accuracy	**72.5% (37 / 51)**	Predicted score within ±15 points of labeled outcome
Mean score delta	**–8.4 points**	Systematic conservative bias — model tends to underestimate stability
Median score delta	**–11 points**	Confirms conservative skew in mid-range predictions
12-month horizon accuracy	**84.3%**	Most defensible external claim window
Mean Brier score (12-month)	**0.169**	Lower is better; naive baseline ≈ 0.250 (~33% improvement)

Key takeaway: at the 12-month horizon — the window most relevant to first-year executive and leadership retention decisions — the engine correctly classified 84.3% of candidates. This performance is approximately 33% better than a naive baseline that assigns equal probability to all candidates.

Study Methodology

Design Principle: No Look-Ahead

Every candidate in this study was scored under simulated pre-hire conditions. The candidate's most recent role — the role whose outcome we wished to predict — was removed from the file before scoring. The engine received only prior job history: roles held before the target position, with start and end dates intact.

This methodology, sometimes called an N-1 temporal holdout, ensures that the signal the engine uses to generate a prediction is identical to what it would see in a real pre-hire context. There is no leakage of outcome information into the feature set.

Outcome Definition

Ground truth for each candidate was the actual duration of employment in the withheld role, measured in months from start to departure or to the observation cut-off. Outcomes were sourced from candidate-provided documentation and verified against employment history in available records.

Accuracy Definition

Score accuracy is reported as the proportion of candidates whose predicted stability score fell within ±15 points of the independent label assigned to that candidate based on actual outcome. The ±15-point window reflects the practical triage resolution the engine is designed to support — it is wide enough to absorb minor calibration variation but narrow enough to enforce meaningful directional accuracy.

Horizon accuracy (3, 6, 12, 24, 36 months) measures the proportion of candidates correctly classified as retained or departed at each time window, using the engine's survival probability output against observed employment outcomes.

Brier Score

The Brier score measures calibration quality of probabilistic forecasts. A score of 0.0 indicates perfect calibration; a naive model that assigns equal probability to all candidates produces a Brier score of approximately 0.25. Lower scores indicate better-calibrated probability estimates.

What the Model Does Not Use

The engine does not use compensation data, interview scores, reference check outcomes, personality assessments, or any information obtained after the offer stage. It relies entirely on structured employment history as available in a standard candidate file.

The study cohort includes candidates across multiple industry sectors, seniority levels, and tenure profiles. Sector-level and archetype-level subgroup slices are tracked separately and will be reported as the live outcome cohort matures.

Validation Results

Horizon-by-Horizon Accuracy

The table below shows classification accuracy at each prediction horizon. The 12-month window is the primary external claim horizon. 24-month and 36-month results are presented for completeness but should be treated as directional signals rather than hard-accuracy claims — survival modeling at longer horizons is inherently subject to greater uncertainty.

Prediction Horizon	Accuracy	Brier Score	Interpretation	Claim Status
3 months	**100.0%**	0.0025	Near-perfect short-term risk detection	**Claim-safe**
6 months	**94.1%**	0.061	High confidence early-tenure risk window	**Claim-safe**
12 months	**84.3%**	0.163	Primary claim horizon — most validated	**Claim-safe**
24 months	43.1%	0.311	Weaker signal; mid-tenure dynamics vary widely	**Directional only**
36 months	60.8%	0.309	Better at 36 than 24 — survival curve behavior	**Directional only**

Note on 24-month accuracy: The 24-month accuracy (43.1%) is lower than the 36-month figure (60.8%). This is a known property of survival modeling — the shape of the survival curve at the 24-month inflection point produces higher variance for borderline candidates. Neither long-horizon figure should be cited as a precision accuracy claim.

Score Accuracy Distribution

The following breakdown shows how predicted stability scores compared to labeled outcomes across the 51-candidate holdout cohort.

Error Band	Candidates	Share	Read
Within ±5 points	**18 / 51**	**35.3%**	High-confidence triage signal
±6–15 points	**19 / 51**	**37.3%**	Within directional accuracy band
±16–25 points	9 / 51	17.6%	Moderate miss — directional use only
Beyond ±25 points	5 / 51	9.8%	Label anomaly or pattern-break case
Total within ±15	**37 / 51**	**72.5%**	Primary accuracy headline

The 5 cases beyond ±25 points include candidates where the N-1 holdout exposed only ambiguous or very thin prior history, making the prediction structurally more difficult — a known ceiling for any model operating on prior job history alone.

Failure Mode Transparency

We report failure modes explicitly. Credible validation evidence discloses where a model is wrong, not only where it is right. The patterns below are stable across the cohort.

Conservative Bias (Mean –8.4 pts)

The engine systematically underestimates stability for candidates with strong prior multi-year histories when the N-1 holdout leaves only a short or thin visible track record. This means the model is more likely to flag a genuinely stable hire than to clear one who is actually a flight risk — a conservative direction of error in the context of executive retention diligence.

Short-Tenure False Positives

Approximately 12% of the cohort (6 of 51) are short-tenure false positives — candidates the model rated as moderately stable who departed within the first year. The most common driver is sparse visible history: when the N-1 holdout removes the only visible recent role, the remaining file provides limited signal. These cases should be surfaced with explicit confidence flagging in the report output.

Long-Tenure False Negatives

Approximately 6% of the cohort (3 of 51) are long-tenure false negatives — candidates the model rated as at-risk who actually stayed 3–5+ years. In every confirmed case, the prior job history available to the model showed genuinely mixed or short-stint patterns that, on the visible evidence alone, justified a conservative read. These are not scorer errors — they are pattern-breaks where the visible history did not predict the actual outcome. Disclosure of this class is appropriate in all client-facing materials.

Known Limitations of This Study

Not yet a live-outcome cohort

Ground truth in this study is derived from historical employment records, not from a prospective observation window on monitored hires. A live-outcome cohort — where candidates are monitored post-hire and outcomes are confirmed in real time — is in progress and will be reported separately as it matures.

Subgroup slices not yet frozen

Accuracy by role archetype (executive vs. individual contributor), industry sector, and seniority band is tracked but not yet published as a frozen artifact. These slices will be added to future validation updates.

No confidence intervals reported

With n=51, formal confidence intervals on the primary accuracy figures would be wide. The reported metrics should be read as point estimates on this cohort, not as population-level precision claims.

Single-model, single-scorer version

This study reflects scorer behavior at commit f0e957e8 (April 3, 2026). Scorer updates are version-controlled and validated against a frozen regression corpus before deployment. Each material version change triggers a new validation cycle.

Interpretation Guide

The Ros Retention Stability Engine is a directional decision-support tool. This guide explains how to use validation evidence appropriately.

What a Stability Score Means

A Stability Score is a ranked signal — it orders candidates by relative retention risk using their prior employment history. Higher scores indicate stronger historical stability patterns; lower scores indicate elevated early-departure risk. The 12-month horizon accuracy of 84.3% applies to classification at that horizon, not to the precision of the underlying score value.

This Engine Is Not a Hire / Do-Not-Hire Tool

Ros produces signal-based diligence, not employment decisions. No score output from this engine should be used as a sole or primary basis for hiring or rejection. Scores are designed to surface questions for interview, reference, and leadership assessment — not to replace those processes. The engine is one input in a broader fiduciary diligence framework.

Recommended Use by Score Band

Score Range	Risk Label	Appropriate Use
75–100	Stability Signal	Standard onboarding diligence. Flag if survival <80% at 12 months.
60–74	Moderate Stability	Validate key tenure transitions and role-change motivations in interview.
45–59	Mixed Signal	Structured retention conversation recommended before close.
30–44	Elevated Risk	Reference-check emphasis on tenure and departure context at each prior role.
Below 30	Critical Risk	Full Logic Trace review recommended. Do not proceed without additional diligence.

Reporting Language

When referencing this study in client materials or investor communications, the following language is accurate and defensible:

> *"In an independent N-1 temporal holdout study across 51 candidates, the Ros Retention Stability Engine correctly classified 84.3% of candidates at the 12-month horizon, with a mean Brier score of 0.169 versus a naive baseline of approximately 0.250. Score accuracy (within ±15 points of labeled outcomes) was 72.5% with a conservative bias of –8.4 points. The engine is a directional risk signal and is not intended as a sole basis for employment decisions."*

Disclosure & Legal Notice

- *This validation study was conducted using historical employment records. Results reflect model performance on this specific 51-candidate cohort and should not be extrapolated to all candidate populations without further validation.*
- *The Ros Retention Stability Engine produces probabilistic, directional signals. It is not a deterministic classification system, and its outputs do not constitute employment decisions, background reports, or consumer reports under applicable law.*
- *Retention outcomes are influenced by factors beyond prior job history, including managerial quality, organizational context, compensation, and macroeconomic conditions. This engine does not capture or model those factors.*
- *Ros produces signal-based diligence and monitoring. It is not designed or intended to be used as the sole or primary basis for any hiring, termination, or promotion decision.*
- *This document is confidential and intended for authorized recipients only. The underlying methodology, algorithm, and scoring system are proprietary and protected. Publication, reproduction, or distribution without written consent from Ros Retention Intelligence is prohibited.*
- *Validation metrics will be updated as the live-outcome cohort matures. The figures in this report reflect the April 2026 frozen snapshot only.*

Ros Retention Intelligence
stabilityengine.ai · ros.stabilityengine.ai
For validation methodology inquiries: audit@stabilityengine.ai

Scorer version: f0e957e8 · Cohort frozen: April 8, 2026 · Document version: 1.0