A/G Scoreboard — Leakage / Bias Audit Memo

Date: April 15, 2026

Scope: Internal audit note covering leakage, bias, and validation structure for the A/G Bank Structural Divergence Scoreboard.

Status: Supporting note — construction integrity and outcome-definition analysis.

---

1. Time Leakage

The validation outcome used throughout the test suite is "4-quarter future deterioration": whether a bank enters Watch or Critical tier in at least one of the next four quarters. This outcome is constructed by iterating forward through the per-bank time series and checking tiers at positions i+1 through i+4 (test_a.py lines 11–18, test_b.py lines 8–15, test_c.py lines 41–48). The observation-quarter features—G, A, MHI, tier classification—are computed solely from the FDIC call report fields reported in that quarter. No forward-looking data enters the feature set.

The temporal features script (temporal_features_split_recovery.py) computes delta_MHI, delta_G, watch_run, and trajectory_class using only same-quarter and prior-quarter information per bank. Peak stress markers (peak_MHI, lowest_G) are computed as expanding windows up to and including the current quarter, not beyond it.

This is a temporal-forward construction. Banks are sorted by CERT and period ordinal before any outcome labeling. There is no shuffled train/test split and no randomization step that could permute time order.

Controlled: Feature computation uses only current-quarter FDIC data. Outcome labeling uses only future tier assignments. No future information enters observation-quarter scores.

Partially addressed: The tier thresholds themselves (G < 0.1 for Critical, MHI > 1.2 for Watch) were set with awareness of the full 34-year panel. They are structural reference points, not fitted parameters, but this distinction could be challenged. Threshold-freeze validation has now been performed (see THRESHOLD_FREEZE_VALIDATION.md): scoring parameters were frozen at production values and evaluated on strictly future data across three temporal splits (pre-2015, pre-2010, pre-2006). MHI decile monotonicity survived intact in all splits. The D10–D1 spread narrowed from 81.4pp (full panel) to 75.5pp in the strictest window (post-2014 only), a modest compression consistent with the shorter evaluation period and lower base rates rather than threshold overfitting. The threshold-awareness concern is addressed but not fully eliminated — the thresholds were still designed with full-panel awareness, and parameter uniqueness (whether other threshold choices would produce equivalent results) has not been tested.

2. Feature Overlap Between A and G

The Ground channel uses four FDIC fields: RBC1AAJ (Tier 1 Leverage Ratio), RBCRWAJ (Total Risk-Based Capital Ratio), NCLNLSR (Noncurrent Loans %), and NTLNLSR (Net Charge-offs %). The Appearance channel uses three: ROA (Return on Assets), EEFFR (Efficiency Ratio), and NIMY (Net

Interest Margin). These are defined in ag_scoring.py lines 21–33 as separate dictionaries under BANKING_CHANNELS.

No variable appears in both channels. This is a deliberate structural constraint: the ratio MHI = A/G is only interpretable as a divergence measure if the numerator and denominator draw from disjoint variable sets. If any feature contributed to both, the ratio would partially cancel itself and lose its diagnostic meaning.

Controlled: Zero variable overlap, verified directly in the channel definitions.

3. Transform and Normalization Artifacts

Each feature is mapped to [0, 1] via a fixed transform with documented reference points. Logistic transforms use domain-motivated centers (e.g., Tier 1 Leverage centered at 5.0%, the FDIC regulatory threshold for "well-capitalized"). Inverse transforms use a scale parameter controlling how aggressively high values of a "bad" metric pull the score toward zero. The inverted logistic for the efficiency ratio is centered at 70%, a conventional boundary between efficient and inefficient banking operations. These reference points are documented, not tuned to maximize any outcome metric.

A production-matched sensitivity harness (sensitivity_analysis.py) has been run perturbing each transform center and slope parameter in both directions, testing alternative G aggregation operators (harmonic mean, weighted min), and measuring the effect on the full validation surface. Under nearby perturbations, the main directional results—within-band lift, monotonic MHI gradient, persistence signal, baseline comparison shape—are broadly preserved. The key structural sensitivity identified is the G aggregation operator: replacing min with harmonic mean shifts tier counts and weakens the bottleneck interpretation without destroying the signal.

Controlled: Transforms use documented reference points, not fitted values. Sensitivity testing covers the parameter neighborhood.

Partially addressed: Full robustness to every plausible alternative transform design is not claimed. The sensitivity harness covers local perturbations around the current design, not a global search over functional forms.

4. Survivorship Bias and Panel Construction

The historical panel is built by loading each quarterly FDIC Financial CSV file (Financial_1992Q1.csv through Financial_2025Q4.csv) via fdic_data.load_all_local(). Each file contains all banks that filed call reports in that quarter. The loader stacks all quarterly files into a single DataFrame with a period column. Every row in every quarterly file is scored and appears in the historical ledger.

Verified empirically:

- The ledger contains exactly the same number of rows per quarter as the source FDIC file (checked for 2008Q4: 8,396; 2010Q1: 8,025; 2025Q4: 4,408).
- All 573 failed banks in the FDIC failure list appear in the historical ledger for their pre-failure quarters (573/573 matched).
- Failed banks have a median of 68 quarters of history in the ledger (range: 10 to 133).

- Zero rows exist at or after the failure quarter for any failed bank.

Banks that fail or close stop appearing in subsequent FDIC quarterly files. The panel includes them for all quarters they existed and excludes them thereafter. This is not survivorship bias in the usual sense—the panel does not retroactively remove failed banks from history. However, the latest quarters are composed entirely of banks that have survived to that date. This is inherent to any panel built from periodic regulatory filings and is standard in banking research.

Controlled: Failed banks are fully represented through the quarters in which they filed reports. No retroactive filtering.

Structural limitation: The declining bank count (14,475 in 1992Q1 to 4,408 in 2025Q4) reflects real consolidation, closure, and failure. Cross-period comparisons should account for this.

5. Internal vs External Validation

Tests A, B, and C are internal temporal-signal tests within the scored framework: they evaluate whether current-quarter A/G structure anticipates later movement into the system's own Watch/Critical tiers and whether that ranking signal remains coherent over time. The failed-bank cross-reference is the main external validation layer currently in place. Taken together, the project now demonstrates internal coherence, temporal signal, and some external alignment, but it has not yet been benchmarked against every possible external event class.

The baseline comparison (baseline_compare.py) compares MHI decile predictiveness against G-only and Texas Ratio proxy deciles using the same outcome definition. This is a within-framework comparison of scoring approaches, not an external benchmark.

Present: Internal coherence tests (A/B/C), one external event layer (failed-bank anchoring), walk-forward external validation against FDIC failure and Texas-ratio distress (see WALKFORWARD_EXTERNAL_VALIDATION.md), expanded benchmark comparison across seven metrics on both internal and external targets (see BENCHMARK_EXPANSION_NOTE.md), within-framework baseline comparison (three metrics on internal target), regime-split validation across nine historical slices. The walk-forward and expanded-benchmark results confirm that G-only is the dominant external predictor; MHI adds internal-tier discrimination but not incremental external failure prediction.

Absent: Benchmarking against CAMELS ratings (confidential), comparison with commercial early-warning systems, or validation against non-failure distress events (enforcement actions, MOU issuances, rating downgrades).

6. Regime Stability

Regime-split validation (regime_stability.py) tested whether the framework's main empirical signatures persist across distinct historical banking environments. Nine slices were evaluated: pre-GFC (1992Q1–2006Q4), GFC stress (2007Q1–2012Q4), post-crisis normal (2013Q1–2019Q4), pandemic (2020Q1–2021Q4), recent softening (2022Q1–2025Q4), and four benchmark temporal splits (pre-2008, crisis-and-after, pre-2020, post-2020).

In all nine tested slices, same-band lift remained above 1.5x, and temporal persistence increased materially with consecutive Watch quarters. MHI decile gradients remained strictly monotonic in most regimes, with reduced rank-order precision in the shorter, lower-stress Pandemic and Recent Softening windows. On internal targets, MHI and G-only remained the two strongest signals across regimes, while the Texas-style proxy was materially weaker and more crisis-dependent. (On external targets, G-only is the dominant predictor and MHI does not add incremental discrimination — see WALKFORWARD_EXTERNAL_VALIDATION.md and BENCHMARK_EXPANSION_NOTE.md.)

The pandemic and recent-softening slices show narrower decile spreads and lose strict monotonicity. This is consistent with their shorter duration (8 and 16 quarters) and lower base rates of deterioration, not evidence of regime-dependent structural failure.

Detailed per-regime results are recorded in output/regime_validation_summary.csv and output/regime_validation_detail.csv.

7. Remaining Open Issues

- **Walk-forward and threshold-freeze validation.** Walk-forward external validation has been completed using three genuinely forward-looking windows (GFC, Post-Crisis, Recent) against FDIC failure and Texas-ratio distress targets (see WALKFORWARD_EXTERNAL_VALIDATION.md). Threshold-freeze validation has also been completed: production scoring parameters were frozen and evaluated on strictly future data across three temporal splits (pre-2015, pre-2010, pre-2006). Monotonicity survived all splits and the qualitative validation story was unchanged. The threshold-awareness concern from Section 1 is now partially addressed rather than fully open. What remains untested is parameter uniqueness — whether materially different threshold choices would produce equivalent validation results.
- **Broader external-outcome benchmarking.** Failed-bank anchoring is one external layer. Additional external events (enforcement actions, deposit outflows, acquisition-under-distress) would strengthen the external case.
- **Alternative aggregator robustness.** A systematic comparison of min, harmonic mean, geometric mean, and weighted min has been completed (see AGGREGATION_OPERATOR_NOTE.md). Min and weighted min preserve Test B monotonicity; geometric and harmonic mean collapse the Watch layer by 90–98% and break monotonicity. External G-only discrimination is comparable across all operators (13–14pp). The remaining open question is whether a broader operator family (e.g., trimmed mean, rank-weighted combinations) or jointly optimized operator-threshold pairings would change the conclusion.
- **Transform functional form.** The current transforms are logistic and inverse. Other reasonable choices (e.g., piecewise linear, log-odds, empirical CDF) have not been tested. The sensitivity harness covers parameter perturbations within the current forms but not alternative forms.
- **Further out-of-sample validation.** The current regime slicing tests the signal within distinct historical windows. Additional designs — such as rolling walk-forward estimation, leave-one-regime-out cross-validation, or testing on newly arriving quarters as they are published — would further strengthen the out-of-sample case.

This memo is an internal audit note. It documents what has been examined, what is controlled, and what remains open. It does not constitute a claim of completeness.