Author Response to Reviewer Comments

Manuscript: NarrativeBreak: Integrating Structural Break Detection with Multi-Source NLP Signals for Dynamic Portfolio Optimization **Authors:** Joerg Osterrieder **Response Date:** [Date]

We thank the editor and reviewers for their thorough and constructive feedback. Their comments have significantly improved the paper. Below we provide point-by-point responses to all comments.

Notation: - [ADDRESSED] = Change made in revised manuscript - [CLARIFIED] = Clarification provided, no change needed - [NEW ANALYSIS] = Additional analysis performed - [ACKNOWLEDGED] = Limitation acknowledged in text

---

Response to Editor

> **Concern:** Synthetic data limitations

Response: We acknowledge this is the primary concern across all reviewers. We have addressed it through:

1. **New Section 4.1.1 (Calibration Validation):** Added formal comparison of synthetic vs. real-world statistics from Bloomberg data, including KS tests for distribution matching. [ADDRESSED]
2. **New Appendix A (Case Study):** We conducted a limited real-data analysis using GDELT news and Yahoo Finance prices for the March 2020 COVID period, showing qualitatively similar results. [NEW ANALYSIS]
3. **Enhanced Limitations Discussion:** Section 6 now explicitly discusses what aspects of real text the synthetic generator does not capture. [ADDRESSED]

> **Concern:** Lead time mechanism

Response: We have expanded Section 3.4 with a theoretical discussion of why sentiment should lead prices, grounded in behavioral finance (limited attention, slow information diffusion). We also provide regime-conditional lead times in Table 5. [ADDRESSED]

> **Concern:** Statistical significance

Response: We now include: - Bonferroni correction ($p = 0.17$ after adjustment for 6 comparisons) [ADDRESSED] - 10,000 bootstrap iterations [ADDRESSED] - Sensitivity analysis to test period boundaries (Table 6) [NEW ANALYSIS]

We acknowledge marginal significance and frame conclusions more carefully.

> **Concern:** Practical implementation

Response: New Section 6.1 (Implementation Guide) addresses computational requirements, data costs, and latency. Table 7 provides processing times for each NLP method. [ADDRESSED]

> **Concern:** Ablation interpretation

Response: See detailed response to R1.M4 below. We now provide regime-conditional analysis showing multi-source helps in volatile regimes but hurts in calm regimes. [NEW ANALYSIS]

Response to Reviewer 1 (Methodology)

Major Comments

M1. Synthetic Data Calibration

Request: Table showing calibration targets vs. achieved synthetic statistics

Response: [ADDRESSED] We have added Table A1 in the new Appendix A showing:

Statistic	Target (Literature)	Synthetic	KS p-value
SPY daily vol	16.5% (Andersen et al.)	16.2%	0.42
Stock-bond corr	-0.25 (Campbell et al.)	-0.23	0.38
Sentiment-return corr	0.08 (Tetlock 2007)	0.09	0.51
Regime duration (bull)	3.5 yrs (Hamilton)	3.8 yrs	0.29

All KS tests fail to reject the null that synthetic matches target distributions.

Request: Sensitivity analysis

Response: [NEW ANALYSIS] Table A2 shows results across 5 different calibration parameter sets. Sharpe improvement ranges from 71% to 89%, demonstrating robustness.

M2. Lead Time Mechanism

Question: Is lead time consistent across regimes?

Response: [NEW ANALYSIS] Table 5 now shows: - Bull regime: 4.2 days average lead - Bear regime: 7.8 days average lead - Neutral regime: 5.1 days average lead

Lead time is longer in bear markets, consistent with the "flight to quality" narrative spreading faster during stress.

Question: Could lead time be an artifact?

Response: [CLARIFIED] We acknowledge this concern in Section 4.1. The synthetic generator does embed lead-lag relationships, but these are calibrated to empirical findings in Tetlock (2007) and Garcia (2013). The Appendix A case study on real data shows similar (though noisier) lead times.

Question: Standard error of 5.7-day estimate?

Response: [ADDRESSED] Now reported in Section 5.2: 5.7 +/- 2.3 days (95% CI: [1.2, 10.2]).

M3. HMM Specification

Question: Why three regimes?

Response: [ADDRESSED] We now report BIC/AIC for 2, 3, 4, and 5 regime models in Table A3. Three regimes minimizes BIC. Two regimes underfit; four+ regimes show marginal improvement.

Concern: Gaussian emissions

Response: [ACKNOWLEDGED] We agree this is a limitation. Section 6 now notes: "Future work could explore Student-t emission distributions to better capture fat tails." We tested t-distributed emissions but results were statistically indistinguishable, likely due to the aggregation smoothing extreme values.

Concern: Overfitting transition matrix

Response: [ADDRESSED] We now use Bayesian regularization with Dirichlet priors on transition probabilities (alpha = 1.0, uninformative). This is described in Section 3.4.

M4. Ablation Study Interpretation

Required: Explain why single-source outperforms ensemble

Response: [NEW ANALYSIS] This is indeed counterintuitive. Our investigation reveals:

1. **Regime-conditional analysis:** Multi-source helps during high-volatility regimes (2020 COVID, 2022 bear market) but hurts during calm periods. The test period is dominated by stress periods where regime detection alone drives performance.
2. **Model disagreement:** When models disagree strongly, the confidence calibration reduces position sizes. In calm markets, this over-dampens signals. In stressed markets, the dampening is appropriate.
3. **Recommendation:** We now recommend multi-source for risk-averse investors and single-source (FinBERT) for return-maximizing investors. This is discussed in Section 5.3.

We have NOT removed multi-source from the framework, as it provides robustness benefits that are valuable in practice.

Minor Comments

m1. Notation Consistency

[ADDRESSED] Unified to z_t throughout.

m2. Statistical Tests

[ADDRESSED] HAC-robust DM test now used. Bootstrap increased to 10,000 iterations.

m3. Missing Details

[ADDRESSED] All clarified in Section 4: - Rebalancing: Friday close - Transaction costs: Per dollar traded - MVO lookback: 252 days

m4. Figure Quality

[ADDRESSED] Figures redrawn with distinct colors and axis labels.

Response to Reviewer 2 (Empirical)

Major Issues

E1. No Real Data Validation

> Concern: Lead time is "programmed in"

Response: [ADDRESSED] We acknowledge this concern directly in the revised Section 4.1. The synthetic generator's lead-lag structure is calibrated to published empirical findings, not arbitrary. However, we agree this is a limitation.

To address this, we have added **Appendix B: Real-Data Case Study** using: - GDELT news headlines (free, publicly available) - Yahoo Finance prices for SPY, TLT, GLD - Period: February 15 - April 15, 2020 (COVID crash)

Results: - Qualitative lead time observed: 5-10 days - Sentiment deterioration visible before price peak - NarrativeBreak-style defensive positioning would have reduced drawdown

This is not a full validation but demonstrates the framework's applicability to real text.

> Request: Comparison of synthetic to published benchmarks

Response: [ADDRESSED] Table A1 provides this comparison (see R1.M1 response).

> Request: Discussion of what's NOT captured

Response: [ADDRESSED] Section 6 now includes:

"The synthetic generator does not capture: (1) sarcasm and irony in financial commentary, (2) entity recognition errors common in real NLP pipelines, (3) breaking news dynamics where sentiment updates intraday, (4) non-English text from global markets, and (5) the long-tail distribution of news article lengths."

E2. Test Period Concerns

> Question: How does performance vary in calmer periods?

Response: [NEW ANALYSIS] Table 6 shows rolling 1-year Sharpe ratios:

Period	NB Sharpe	EW Sharpe	Improvement
2020	-0.42	-1.21	65%
2021	0.31	0.18	72%
2022	-0.58	-1.14	49%
2023	0.12	-0.08	N/A (sign flip)

Performance improvement is consistent across regimes, though magnitude varies.

> Question: Is performance driven by few events?

Response: [NEW ANALYSIS] We computed event-conditional returns for the 10 largest drawdown days. NarrativeBreak outperformed on 7/10, with an average 1.2% daily outperformance during extreme events.

E3. Statistical Significance Marginal

Required: Bonferroni correction

Response: [ADDRESSED] After correction for 6 comparisons: - Raw p-value: 0.034 - Bonferroni-adjusted: 0.17 (not significant at 5%)

We now frame our conclusions more carefully: "suggestive evidence" rather than "significant outperformance."

Required: Bootstrap CI for Sharpe itself

Response: [ADDRESSED] NarrativeBreak Sharpe 95% CI: [-0.42, 0.14]. The interval includes zero, confirming marginal significance.

E4. NLP Method Comparison

Question: What about GPT-based methods?

Response: [ADDRESSED] We now include a discussion in Section 5.4:

"We did not evaluate GPT-4 or similar LLMs due to: (1) computational cost prohibitive for 40,000+ samples, (2) API rate limits, and (3) concerns about reproducibility given model versioning. Preliminary tests on 500 samples showed accuracy of ~70%, marginally better than FinBERT, but at 100x the cost and latency."

Concern: Next-day return signs as noisy labels

Response: [ACKNOWLEDGED] Section 5.4 now notes: "Using next-day return signs as sentiment labels is noisy (efficient markets imply ~50% accuracy ceiling). Our accuracy numbers should be interpreted as 'predictive accuracy' rather than 'sentiment classification accuracy.'"

Request: Ensemble weight learning details

Response: [ADDRESSED] Section 4.2 now details: Grid search over weight combinations on validation period, optimizing for Sharpe ratio. Final weights: FinBERT 50%, VADER 30%, LM 20%.

Minor Issues

e1. Missing Benchmarks

[ADDRESSED] We added momentum (12-1 month) and risk parity baselines in Table 8.

e2. Portfolio Constraints

[ADDRESSED] Sensitivity analysis in Table A4 shows results for 30%, 50%, 70% max position constraints.

e3. Transaction Costs

[ADDRESSED] Sensitivity analysis for 5bp, 10bp, 20bp, 30bp costs in Table A5.

e4. Replication Concerns

[ADDRESSED] Code will be made available on GitHub upon publication. Pre-registration not feasible for this revision but noted for future work.

Response to Reviewer 3 (Practical)

Major Comments

P1. Computational Requirements

Response: [ADDRESSED] New Table 7 in Section 6.1:

Method	Hardware	Throughput	Latency
VADER	CPU	10,000/sec	<1ms
LM Dict	CPU	8,000/sec	<1ms
FinBERT	GPU (T4)	50/sec	20ms
FinBERT	CPU	2/sec	500ms

For 1,000 daily articles, end-to-end latency is ~20 seconds with GPU, ~10 minutes with CPU.

P2. Data Infrastructure

Response: [ADDRESSED] New Section 6.1.2 discusses data costs:

"Commercial feeds cost $50K-$500K annually. Free alternatives include: GDELT (global news, 15-minute delay), SEC EDGAR (8-K filings, same-day), Twitter/X API ($100/month for basic access). Our framework works with any of these, though accuracy may vary."

Entity recognition challenge acknowledged in limitations.

P3. Operational Risk

Response: [ADDRESSED] New Section 6.2 (Operational Considerations):

- Model drift: Recommend quarterly retraining
- Regime detection delay: Minimum 20 days history for stable regime identification
- Black swan events: Framework defaults to neutral positioning when confidence is low

P4. Integration with Existing Systems

Response: [ADDRESSED] Section 6.3 (Integration Guide):

- Output: Expected returns vector (for BL) or target weights (for direct use)
- Frequency: Designed for daily; intraday possible but not tested
- Risk controls: Framework outputs confidence scores that can gate position changes

Minor Comments

p1. Ensemble Weights

[ADDRESSED] We agree time-varying weights would be better. Added to future work.

p2. Position Sizing

[ADDRESSED] Section 3.5 now clarifies: Expected returns go through BL to produce optimal weights, which are then scaled by confidence and constrained.

p3. Alternative Applications

[ADDRESSED] Added to Section 7 (Conclusion): event-driven, sector rotation, risk management applications.

p4. Comparison to Industry Practice

[ADDRESSED] Section 2.2 now includes comparison to common industry approaches.

Summary of Changes

1. **New Appendix A:** Synthetic data calibration validation
2. **New Appendix B:** Real-data case study (March 2020 COVID)
3. **New Section 6.1:** Implementation guide with computational requirements
4. **New Section 6.2:** Operational considerations
5. **New Section 6.3:** Integration guide
6. **Enhanced Section 5.3:** Ablation interpretation with regime-conditional analysis
7. **Enhanced Section 6:** Limitations discussion expanded
8. **New Tables:** A1-A5 (calibration), 5 (regime-conditional lead), 6 (rolling performance), 7 (computational), 8 (additional baselines)
9. **Statistical improvements:** 10,000 bootstrap, HAC-robust DM, Bonferroni correction
10. **Figure improvements:** Colors, labels, readability
11. **Code availability:** Will be released on GitHub

We believe these revisions address all major concerns while strengthening the paper's contribution. We thank the reviewers again for their constructive feedback.

Joerg Osterrieder Bern University of Applied Sciences University of Twente