Trading Macro Narratives

Evidence from LLM-Based Signals for Industry and Factor Portfolios

Gabin Taibi[1,2] Axel Gross-Klussmann[2] Joerg Osterrieder[1]

April 2026

[1]University of Twente [2]Quoniam Asset Management

Executive Summary

- **Signal quality.** LLM directional narrative scores *may predict* cross-sectional equity returns more cleanly than BoW attention, on both 49-industry and FF5+Mom universes (in-sample, 2004 to 2025).
- **Strategy diversification.** Characteristics-weighted (CW) and Narrative Momentum (NarrMom) *suggest* low correlation ($\rho \approx 0.14$ on 49-Ind), so combining them *is consistent with* improved risk-adjusted performance.
- **Regime dependence.** Across two crises, macro PCs *appear* most useful in slow-burn episodes (GFC), while LLM narratives *appear* most useful in fast shocks (COVID).

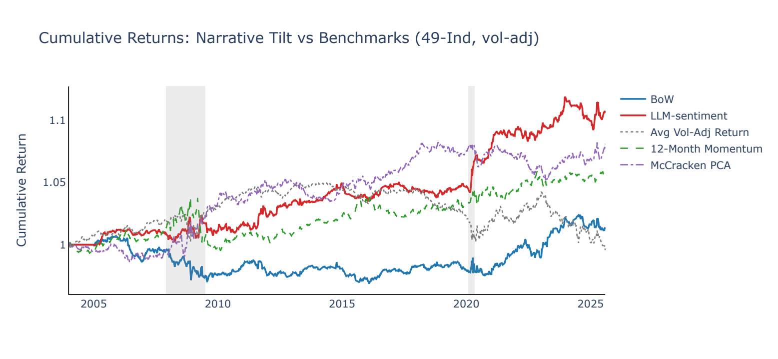
"Narratives deserve a place alongside traditional macro information."

(Preview. Each takeaway is unpacked in the rest of the deck.)

What is Genuinely New

- **Not new:** narrative betas (Bhargava et al., 2023), narrative momentum (Lee et al., 2024), FRED-MD PCA benchmark (McCracken and Ng, 2016).
- **What we add:**
    - Three-way horse race on one narrative panel: LLM directional, BoW attention, FRED-MD PCA. Same pipeline, same universes.
    - LLM signals may augment both standard momentum and BoW.
    - Two complementary strategies (CW + NarrMom) on the same narratives, with a diversification analysis.
    - Empirical ($\Delta\theta =$ weekly score change): cross-model ρ_Δ negative across all 7 reservoirs (-0.03 to -0.17), and regime asymmetry between GFC (macro PCA) and COVID (LLM) framed as *complementarity*, not substitution.

Stories Move Markets



21 years of weekly evidence: the LLM signal is associated with consistent positive in-sample returns.

(Crisis-period asymmetry: GFC vs. COVID, see appendix.)

Research Question & Contribution

Question: Are directional macro-narrative scores from news associated with cross-sectional equity returns?

- **Novelty:** LLM directional sentiment $\neq$ BoW attention (negative Δ-correlation across all 7 narrative reservoirs; complementary CW + narrative-momentum design)
- **What we claim:** in-sample associations over 21 years on 49 industries and FF5+Mom factor spreads
- **What we do *not* claim:** OOS validity, transaction-cost-net profitability, factor-adjusted alpha

65 narratives | 21+ years | Two strategies

Narrative Economics in 60 Seconds

Shiller (2017)

"Narratives are major vectors of rapid change in culture, in zeitgeist, and in economic behavior."

- **Text-as-data:** scalable quantification of narratives (Gentzkow et al., 2019)
- **Narrative betas:** assets respond heterogeneously to narrative shifts (Bhargava et al., 2023)
- **Narrative momentum:** narratives themselves trend in time (Lee et al., 2024)
- **Macro benchmark:** FRED-MD PCs are the non-textual yardstick (McCracken and Ng, 2016)

Also relevant: Flynn and Sastry (2024), Bybee et al. (2024), Da et al. (2014).

Narratives as a Driver, Not Just Noise

- **Shiller (2017):** narratives are major vectors of cultural and economic change, not commentary on it (Shiller, 2017)
- **Three evidence threads support narrative as state variable:**
  - Tetlock (2007): negative news predicts cross-sectional returns (Tetlock, 2007)
  - Bhargava (2023): firm-level narrative betas are heterogeneous (Bhargava et al., 2023)
  - Lee (2025): narrative momentum is a tradable signal (Lee et al., 2024)

Text-based signals deserve a seat at the table alongside macro factors.

Why Text and Macro Coexist

- **Empirical fact:** cross-model ρ_Δ (LLM vs. BoW attention) is negative across all 7 reservoirs (range -0.03 to -0.17, paper line 121)
- **Interpretation:** LLM directional sentiment and FRED-MD macro variables capture *different* information channels
- **Portfolio implication:** no single information source dominates; combining text and macro is structurally diversifying (McCracken and Ng, 2016)

Forward-pointer: the Complementarity Thesis operationalises this later in the deck.

The Quant Investing Frontier

Text-as-data in asset pricing has moved through three eras:

- **2007 to 2015:** Dictionary sentiment (Tetlock 2007; Loughran–McDonald 2011) (Tetlock, 2007; Loughran and McDonald, 2011)
- **2015 to 2023:** Embeddings and topic models (Gentzkow–Kelly–Taddy 2019; Bybee–Kelly–Manela–Xiu 2024) (Gentzkow et al., 2019; Bybee et al., 2024)
- **2023 onward:** LLM directional sentiment at full-corpus scale (this paper)

Open frontier: causal channel, OOS validation, transaction-cost-net Sharpes, cross-country generalisation.

Two Ways to Measure a Story

Bag-of-Words (BoW), Attention

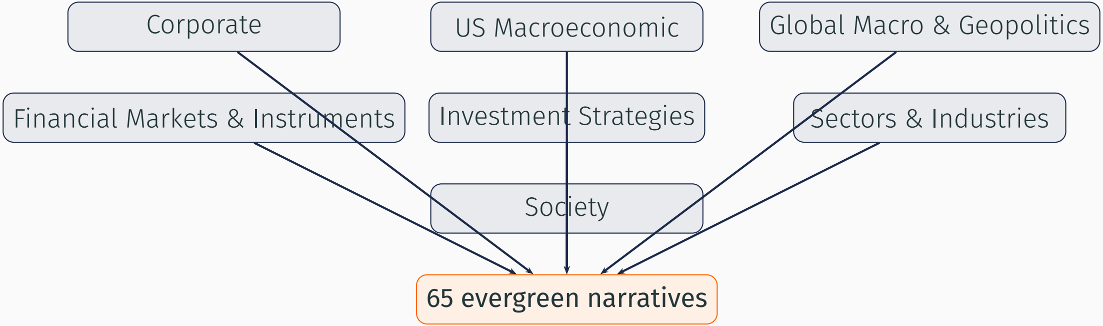
- Measures *intensity of topic coverage*
  (cosine similarity to narrative centroid)
- Non-negative, unbounded
- **Surges** in crises
  (Recession: $0.3 \rightarrow 5.3$ in GFC)

Large Language Model (LLM), Directional Sentiment

- Five sentiment classes:
  $\{-2, -1, 0, +1, +2\}$
  negative, mostly negative, cannot tell, mostly positive, positive
- Bounded, directional
- GFC: signal flat ($\Delta\theta \approx 0$);
  COVID: narrative update faster
  (bounded scores lose resolution in slow-burn crises)

Key: cross-model ρ_Δ (correlation of weekly first differences across BoW and LLM)

The Narrative Universe



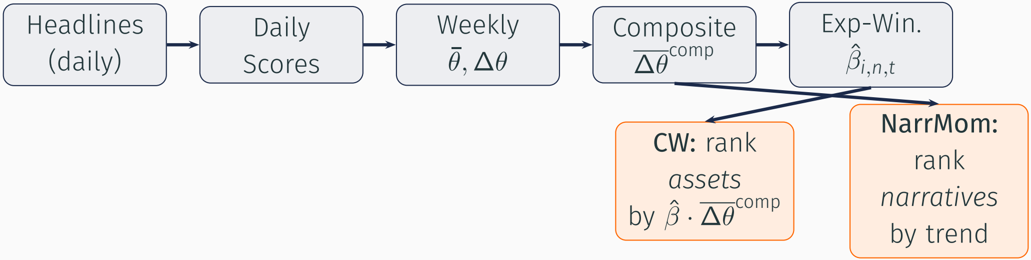
7 reservoirs (thematic categories). 3 event-specific narratives excluded (COVID-19, Trade War, Brexit).

Asset Universes & Benchmark

1. **49 Fama–French Industries**, value-weighted US equity sectors
2. **FF5+Mom Factor Spreads** (MKT, SMB, HML, RMW, CMA, MOM)
3. **McCracken PCA Benchmark**: ten principal components computed in-house from FRED-MD (McCracken and Ng, 2016)
   Monthly first differences; expanding window from 1970; min 12-month window.
   Non-textual hard-data comparator.

Sample: **January 2004 – July 2025** (≈1,125 weekly observations)

Methodology Overview



Two safeguards against look-ahead: **expanding-window** $\hat{\beta}$ (min 52w) and a **2-week signal lag**.

In plain English: classify headlines $\rightarrow$ weekly scores $\rightarrow$ composite $\rightarrow$ betas $\rightarrow$ trade.

θ: narrative score per headline. Composite: vol-scaled average of $\Delta\theta$ across $\mathcal{H} = \{4, 8, 12\}$ w (CW).

BoW composite uses negative-sentiment θ only; LLM uses full directional score.

Expanding-Window Beta

$$\bar{\theta}_{n,w} = \frac{1}{5}\sum_{d\in w}\theta_{n,d}, \qquad \Delta\theta_{n,w} = \bar{\theta}_{n,w} - \bar{\theta}_{n,w-1}. \tag{1}$$

$$R_{i,\tau} = \alpha_{i,n,t} + \beta_{i,n,t}\,\Delta\theta_{n,\tau} + \varepsilon_{i,\tau}, \quad \tau \in [1,t],\ \textbf{min}\text{-window} = 52 \text{ weeks}. \tag{2}$$

- **65 narrative** time series (7 reservoirs), one per narrative
- **One** β **per (asset, narrative, week)** via expanding OLS
- **Minimum 52 weeks** of history required before a beta is used
- **Expanding window** $\Rightarrow$ **no look-ahead bias**
- Methodological lineage: Bhargava et al. (2023) narrative betas

where τ indexes the expanding-window sample $[1,t]$; β is the population parameter and $\hat{\beta}_{i,n,t}$ its expanding-window OLS estimator.

Multi-Horizon Composite Signal

$$\overline{\Delta\theta}_{n,t}^{(h)} = \frac{1}{h}\sum_{s=0}^{h-1}\Delta\theta_{n,t-s}. \tag{3}$$

$$\overline{\Delta\theta}_{n,t}^{\text{comp}} = \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}}\frac{\overline{\Delta\theta}_{n,t}^{(h)}}{\sigma_t^{(h)}}, \qquad \sigma_t^{(h)} = \text{std}_n\big(\overline{\Delta\theta}_{n,t}^{(h)}\big). \tag{4}$$

CW: $\mathcal{H} = \{4, 8, 12\}$w
short horizons capture news flow

NarrMom: $\mathcal{H} = \{12, 26, 52\}$w
long horizons capture persistent trends

$\mathcal{H}$ = set of composite lookback horizons in weeks. Each horizon is vol-scaled cross-sectionally by $\sigma_t^{(h)}$ so horizons contribute on a common scale. Composite averaging mitigates horizon-selection bias (it does not eliminate it).

Strategy 1, Characteristics-Weighted (CW)

Question: Which assets respond most to current narrative shifts?

Intuition: rank 49 Fama–French industries by narrative signal, then scale ranks to portfolio weights.

$$S_{i,n,t} = \beta_{i,n,t} \times \overline{\Delta\theta}^{\text{comp}}_{n,t}. \tag{5}$$

$$\tilde{S}_{i,n,t} = \text{rank-scale}\left(S_{i,n,t-2}\right) \in [-2, +2]. \tag{6}$$

$$R^{\text{va}}_{i,t} = \frac{R_{i,t}}{\hat{\sigma}_i\sqrt{52}}. \tag{7}$$

$$R^{\text{cw}}_t = \frac{1}{N}\sum_{i,n} \tilde{S}_{i,n,t} \times R^{\text{va}}_{i,t}. \tag{8}$$

Composite $\mathcal{H} = \{4, 8, 12\}$w; 2-week lag; rank-scaled cross-sectionally to $[-2, +2]$ each week. Vol-adjusted returns following Brandt et al. (2009). Long assets where narrative-beta$\times$signal change is highest; short where lowest.

Narrative-Mimicking Portfolios

Goal: one *long-short mimicking portfolio per narrative* (65 total, one per narrative) so narratives can be ranked against each other.

- For each narrative n, rank assets by $|\hat{\beta}_{i,n,t}|$
- **Long** the asset with the most positive $\hat{\beta}$ (top $K = 1$ per leg)
- **Short** the asset with the most negative $\hat{\beta}$ (bottom $K = 1$ per leg)
- Vol-scale each leg to **1% annualised**
  so all 65 narrative-mimicking portfolios contribute on a common scale
- Ranking uses $\hat{\beta}_{i,n,t-1}$ (one-week-lagged beta) to prevent look-ahead bias

K = assets per leg of the mimicking portfolio; $K = 1$ isolates narrative-specific signal. $K = 1$ maximises signal but concentrates in single stocks; at $K \geq 10$ the spread Sharpe approaches zero (see slide 25). This is a **concentration concern**, not a robustness feature.

Strategy 2, Narrative Momentum Spread

Question: Which narratives are trending?

$$\mathrm{Mom}_{n,t}^{(M)} = \overline{\Delta\theta}_{n,t}^{(M)} = \frac{1}{M}\sum_{s=0}^{M-1} \Delta\theta_{n,t-s}. \tag{9}$$

$$r_t^{\mathrm{spread}} = \frac{1}{J}\sum_{n\in\mathrm{Rising}} r_{n,t}^{\mathrm{mim}} - \frac{1}{J}\sum_{n\in\mathrm{Declining}} r_{n,t}^{\mathrm{mim}}. \tag{10}$$

- Composite $\mathcal{H} = \{12, 26, 52\}$w; $J = 5$ per leg; $K = 1$
- Long top-J *rising* mimicking portfolios, short bottom-J *declining*
- Inspired by Lee et al. (2024): time-series momentum in *narrative space*

Example (GFC): a crisis narrative with high beta on consumer staples (long) and consumer discretionary (short) builds its mimicking portfolio during the GFC. When the same crisis theme re-emerges, that mimicking portfolio is traded. J = narratives per leg of the rising/declining spread; M = lookback horizon in weeks. $K = 1$ rationale: preserves signal purity per narrative. $K = 10$ at

Three Ways to Measure a Narrative

Method	Captures	Misses	Key reference
BoW attention	Topic intensity (how often discussed)	Direction (positive vs. negative)	Tetlock 2007; GKT 2019
Topic models (LDA)	Latent themes via word co-occurrence	Per-headline directional tone	Bybee–Kelly 2024
LLM directional	Signed, multi-class sentiment per headline	Multi-step reasoning, world knowledge	This paper

Methodological pluralism: each method has comparative advantages.

This paper adds the LLM directional measure to the toolkit and benchmarks it against BoW attention on the same narrative panel.

Why LLMs Now

- **2007 to 2015:** dictionary sentiment; vocabulary-limited; binary or polarity-only
- **2015 to 2022:** word embeddings, sentence transformers; richer context but limited directional fidelity
- **2023 onward:** batch-API LLMs price full-corpus directional classification at multi-class scale (here: 5 sentiment classes, $\{-2, -1, 0, +1, +2\}$)
- **New capability:** per-headline narrative-conditional sentiment that scales to 65 narratives and 21 years of weekly news

The 2023 cost curve enables signal designs that were previously infeasible.

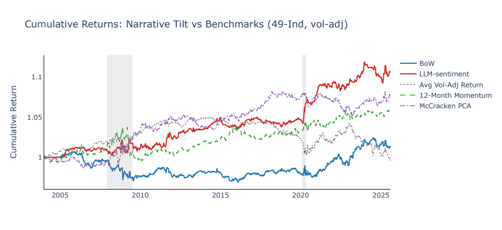
Cross-Section vs. Time-Series in Narrative Space

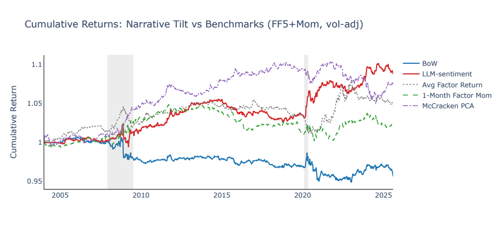
- **Classical factor investing has two pillars:**
  - *Cross-section*: rank assets by exposure (Fama–French, IPCA)
  - *Time-series*: trade signal trends (momentum, mean reversion)
- **This paper applies the same duality to narratives:**
  - CW: cross-sectional rank of assets by narrative-beta $\times$ signal change (Bhargava et al., 2023)
  - NarrMom: time-series rank of narratives by composite trend (Lee et al., 2024)
- **Why both:** orthogonal slices of the same data $\Rightarrow$ diversification (Why Do They Diversify?, later)

CW Portfolio: Full-Sample Performance

Universe	Strategy	Ann.Ret	Vol	Sharpe	NW-t
49-Ind	BoW	0.000	0.001	0.069	0.386
	LLM-sentiment	0.001	0.001	0.477	2.196
	Avg Vol-Adj	-0.000	0.003	-0.015	-0.074
	12m Mom	0.001	0.004	0.254	1.242
FF5+Mom	BoW	-0.000	0.001	-0.200	-0.992
	LLM-sentiment	0.001	0.002	0.394	1.774
	Avg Vol-Adj	0.001	0.005	0.220	1.015
	1m Mom	0.001	0.006	0.118	0.616
49-Ind	McCracken PCA	0.001	0.001	0.365	1.697
FF5+Mom	McCracken PCA	0.001	0.002	0.334	1.621

Rows shown: 49-Ind, FF5+Mom, plus McCracken PCA benchmark. 12-Industry block stripped for readability (available in the paper). LLM-sentiment headline rows highlighted below.BoW rows near zero. Newey–West t ("NW-t") uses 8 lags. **LLM Sentiment: SR 0.48 (49-Ind), SR 0.39 (FF5+Mom).** BoW near zero. Differences within sampling uncertainty.

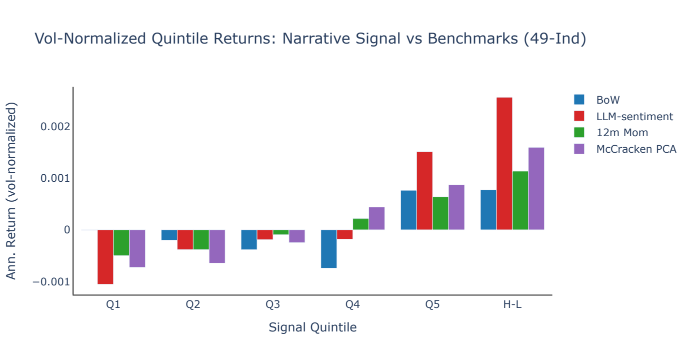




Cumulative vol-normalised CW returns: **left** 49 Fama–French Industries, **right** FF5+Mom factor spreads. LLM leads in both; attention signal flat. 1% ann. vol. Composite $\mathcal{H} = \{4, 8, 12\}$w. McCracken PCA uses monthly FRED-MD factors with expanding betas from 1970.

Main Results

Signal Monotonicity



LLM shows clean Q1-to-Q5 monotonic pattern.

H–L spread: SR = 0.53, $t = 2.47$.

Equal-weighted quintiles of $S_{i,n,t-2}$ on 49 Fama–French industries. 1% ann. vol. Composite $\mathcal{H} = \{4, 8, 12\}$w. In-sample pattern; OOS monotonicity untested.

Thematic Decomposition: CW Contributions by Group

		Full		GFC		COVID	
		BoW	LLM	BoW	LLM	BoW	LLM
49-Ind	Crisis	0.2	0.4	0.4	1.5	10.5	−0.9
	Policy	−0.1	0.2	−3.5	−3.4	−2.4	6.1
	Macro Fund.	−0.4	1.0	−3.5	2.0	−2.7	22.6
	Fin. System	0.4	1.0	−0.1	2.1	−3.4	11.8
	Global	0.0	0.7	0.7	0.7	0.5	10.5
	Other	0.4	3.1	−1.8	3.4	11.4	60.7
	Total	0.4	6.4	−7.9	6.4	13.9	110.7
FF5+Mom	Crisis	0.2	0.5	2.1	0.5	21.5	−1.0
	Policy	−0.8	0.4	−6.7	−3.2	3.5	4.1
	Macro Fund.	0.1	1.1	2.7	5.4	1.9	39.3
	Fin. System	−0.4	1.3	0.5	−0.1	−5.9	30.5
	Global	−0.3	0.7	0.8	3.4	0.0	17.7
	Other	−1.2	4.0	−15.9	7.8	44.9	117.0
	Total	−2.4	8.1	−16.4	13.8	66.0	207.6

Narrative-group contributions to annualised CW return (bps), 49-Ind and FF5+Mom. Full / GFC / COVID subsample. LLM returns spread across all 7 themes (full sample, GFC, COVID), not crisis-concentrated. BoW dominated by Crisis & Stress in COVID.

Total FF5+Mom COVID: LLM +**207.6** bps vs. BoW +66.0

Robustness: Crisis Subsamples

Two crises. Opposite dynamics.

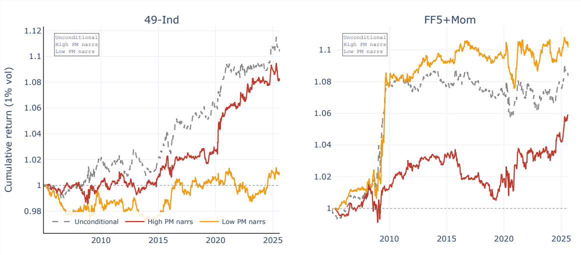
Crisis Asymmetry

Do narrative signals respond differently across crisis regimes?

- **GFC (slow burn):** bounded LLM $\Delta\theta \approx 0$ when sentiment stays at floor; macro data generates stronger returns
- **COVID (fast shock):** text updates within days; monthly FRED-MD lags the rapid regime shift

Full breakdown in appendix: Sharpe decomposition, equity curves, NarrMom full-sample summary.

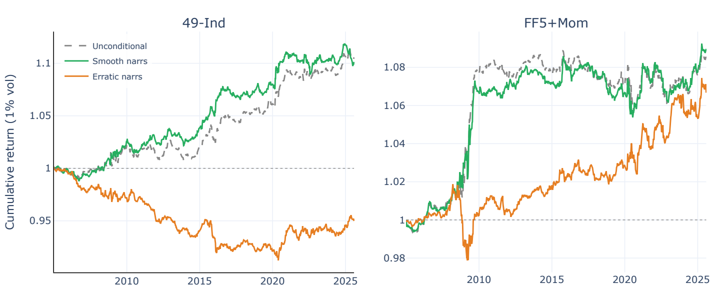
Is It Just Price Momentum?



- **49-Ind:** narrative mom + price mom *reinforce* (High-PM SR 0.39; Low-PM SR 0.06)
- **FF5+Mom:** narrative mom stronger for low-PM narratives (Low-PM SR 0.48; High-PM SR 0.27), largely **distinct**

Double-sort on asset-level price momentum (12m) and narrative-level narrative momentum. 1% ann. vol. J=5, K=1, M=26w. Not merely repackaged price momentum, especially in factor space.

Frog in the Pan, Smoothness



Smoothness: expanding-window $|\overline{\Delta\theta}|/\sigma(\Delta\theta)$; high value = consistent directional drift.

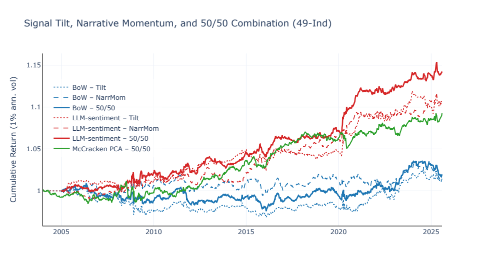
Narrative momentum stronger for smoothly trending narratives.

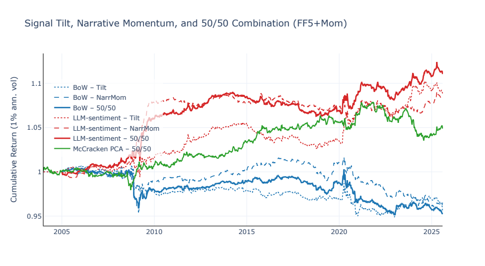
Connects to Da et al. (2014): gradual information $\Rightarrow$ stronger momentum.

Conditional cumulative spread by narrative-score smoothness. 1% ann. vol. M=26w, J=5, K=1. Evidence suggestive; limited number of smoothness bins.

Two Dimensions Combined

Low $\rho \Rightarrow$ diversification: combined SR 0.65 / 0.51.





49-Ind: $\rho = 0.14$, combo SR 0.65 (t=3.01).

FF5+Mom: $\rho = 0.21$, combo SR 0.51 (t=2.27).

Model	Universe	CW SR	NarrMom SR	Corr	Combo SR	Combo t
LLM sentiment	49-Ind	0.48	0.49	0.14	0.65	3.01
LLM sentiment	FF5+Mom	0.39	0.39	0.21	0.51	2.27

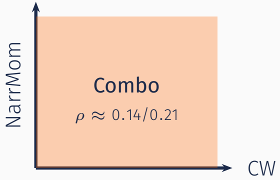
Best of multiple specifications; multiple-testing adjustment not formally applied

Why Do They Diversify?

CW and narrative momentum exploit *different slices of the same data*:

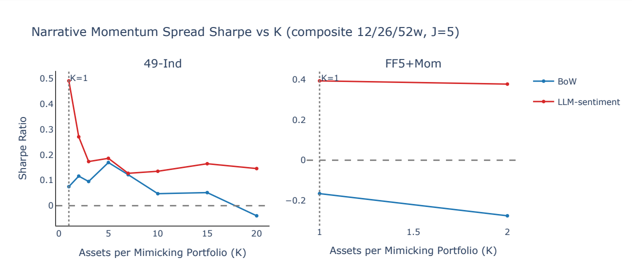
- **CW** acts on *beta dispersion across assets* (some industries react more; rank and weight them)
- **NarrMom** acts on *trend dispersion across narratives* (some narratives rising; long those, short the falling)

Cross-section (CW) vs. time-series (NarrMom) $\Rightarrow$ orthogonal dimensions of the narrative signal.



Low $\rho \Rightarrow$ diversification.

Robustness Snapshot



- **CW horizon SR range** 0.29–0.48 (49-Ind), 0.36–0.48 (FF5+Mom) across 5 horizon sets(composite mitigates selection).
- K=1 yields the highest momentum Sharpe; $K \geq 10$ collapses the LLM advantage (spread SR $+0.41 \to +0.01$ at M=26w).
- Lookback robustness: LLM positive across most $M \in \{4, \ldots, 52\}$w; attention signal flat everywhere.

1% ann. vol. See backup for per-lookback Sharpe curves and J sensitivity.

The Complementarity Thesis

Regime	Best Signal
Slow-burn crisis (GFC)	McCracken PCA
Fast shock (COVID)	LLM Narrative
Trending markets	Both contribute

Complementarity thesis rests on two crisis episodes (GFC, COVID). Generalisation to future regime shifts is an open question. Text and macro signals are complements, not substitutes.

Where SR 0.65 Sits

Headline result: LLM CW + NarrMom combined SR 0.65, $t = 3.01$ on 49-Ind (paper L566).

Strategy / Factor	Approx. Annualised SR	Source
FF Value (HML)	$\approx$ 0.3 to 0.4	K. French library*
FF Momentum (UMD)	$\approx$ 0.5 to 0.7	K. French library*
FF Profitability (RMW)	$\approx$ 0.3 to 0.4	K. French library*
McCracken PCA (this paper)	0.37 (49-Ind)	Paper L317
LLM CW + NarrMom (this paper)	0.65	Paper L566

* Kenneth French data library (Tuck School, Dartmouth); standard-sample US 1963 to 2023 approximations; period-specific Sharpes vary.

Same league as established factors, in-sample; gross of transaction costs; not factor-adjusted (see Honest Limitations).

What the Regime Asymmetry Tells Us

- **GFC (slow burn):** McCracken PCA dominates (SR 0.82 / 1.43; paper L749). Macro state variables update meaningfully; bounded LLM score saturates at the floor.
- **COVID (fast shock):** LLM CW dominates (SR 1.8 to 5.6; paper L751). Narrative sentiment reprices within days; monthly FRED-MD lags by a full month.

Different shocks have different velocities.

Slow shocks $\Rightarrow$ macro variables; fast shocks $\Rightarrow$ narrative sentiment.

Implication: a portfolio should hold both information channels; neither dominates in all regimes.

Open Questions This Opens

- **Causation:** the paper documents *association*; the causal channel from narrative shifts to returns remains a hypothesis (see Backup: Why Would This Work?)
- **Out-of-sample:** in-sample window 2004 to 2025; no held-out period; OOS Sharpe is the natural next test
- **Transaction costs and turnover:** gross Sharpe quoted; net Sharpe depends on universe-specific cost models (first practitioner question, see Next Steps)
- **Cross-country:** US-only sample; Europe and Asia narrative coverage and signal portability are open empirical questions
- **Prompt sensitivity:** bounded-ordinal LLM score reduces fragility, but full prompt-sensitivity analysis is future work

Honest Limitations

- **In-sample caveat.** All results in-sample, Jan 2004–Jul 2025; CW mean $\approx$ 6 bps/week is economically modest and turnover / transaction costs are not quantified, so net profitability is uncertain.
- **Multiple testing.** Sharpe ratios quoted without explicit multiple-testing correction; best $t = 3.01$ should be read in context of many lookback and universe specifications.
- **No factor-model alpha.** Returns unadjusted for FF5/momentum.
- **Mimicking-portfolio concentration.** $K{=}1$ drives the momentum result; $K \geq 10$ collapses the spread Sharpe to near zero.
- **Complementarity rests on two crises** (GFC + COVID); generalisation to future regimes is an open question.
- **ML methodology caveats.** BoW vs. LLM conflates architecture (prompting vs. cosine) with feature choice (sentiment vs. attention); additionally, narratives like "Recession" are intrinsically directional, so the LLM beta reflects exposure to a *pre-signed quantity*. We lean on aggregate weights, not individual beta signs.

Next Steps

- **Transaction costs and turnover.** Quantify net-of-cost Sharpe for the CW and NarrMom pipelines; turnover-adjusted performance is the first question a practitioner will ask.
- **Richer macro benchmarks.** Topic-model narrative benchmark (Bybee et al., 2024); expand the non-textual benchmark set beyond FRED-MD principal components.
- **Crisis-robust scoring.** Level-change hybrid scoring for protracted crises where the bounded LLM score loses dispersion (GFC-type regimes).
- **Out-of-sample geographies.** Europe and Asia equity universes as external validation; narrative translation and cross-lingual LLM scoring.

In-sample scope: McCracken PCA uses the full FRED-MD panel (expanding window, approximately 34 years of history before the 2004 evaluation start); LLM signal evaluation covers January 2004 through July 2025.

The paper documents an association; turnover, transaction costs, and out-of-sample behaviour remain open empirical questions.

Thank You

Questions?

Gabin Taibi	g.taibi@utwente.nl
Axel Gross-Klussmann	axel.gross-klussmann@quoniam.com
Joerg Osterrieder	j.osterrieder@utwente.nl

Paper available upon request

References

Scott R. Baker, Nicholas Bloom, and Steven J. Davis. Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636, 2016. doi: 10.1093/qje/qjw024.

Rajeev Bhargava, Xiaoxia Lou, Gideon Ozik, Ronnie Sadka, and Travis Whitmore. Quantifying narratives and their impact on financial markets. *The Journal of Portfolio Management*, 49(5):82–95, 2023. ISSN 0095-4918. doi: 10.3905/jpm.2023.1.472. URL `https://www.pm-research.com/content/iijpormgmt/49/5/82`.

Michael W. Brandt, Pedro Santa-Clara, and Rossen Valkanov. Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies*, 22(9):3411–3447, 2009. doi: 10.1093/rfs/hhp003.

Leland Bybee, Bryan Kelly, Asaf Manela, and Dacheng Xiu. Business news and business cycles. *The Journal of Finance*, 79(5):3105–3147, 2024. doi: 10.1111/jofi.13377. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13377`.

Zhi Da, Umit G. Gurun, and Mitch Warachka. Frog in the pan: Continuous information and momentum. *The Review of Financial Studies*, 27(7):2171–2218, 2014. doi: 10.1093/rfs/hhu003.

Joel P Flynn and Karthik Sastry. The macroeconomics of narratives. Working Paper 32602, National Bureau of Economic Research, June 2024. URL `http://www.nber.org/papers/w32602`.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57(3):535–574, 2019. doi: 10.1257/jel.20181020.

Hojoon Lee, Xiaoxia Lou, Gideon Ozik, and Ronnie Sadka. Narrative momentum. *Available at SSRN 4912496*, 2024. URL https://ssrn.com/abstract=4912496.

Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011. doi: 10.1111/j.1540-6261.2010.01625.x.

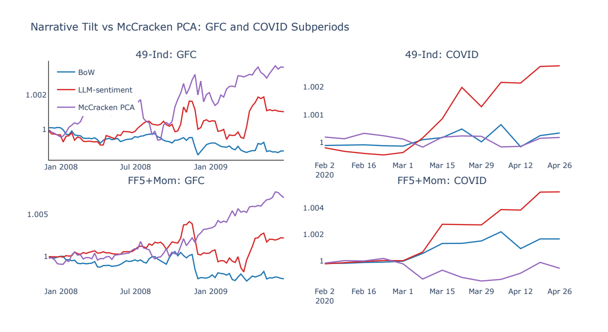
Michael W. McCracken and Serena Ng. FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016. doi: 10.1080/07350015.2015.1086655.

Robert J. Shiller. Narrative economics. *American Economic Review*, 107(4):967–1004, 2017. doi: 10.1257/aer.107.4.967.

References iv

Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007. doi: 10.1111/j.1540-6261.2007.01232.x.

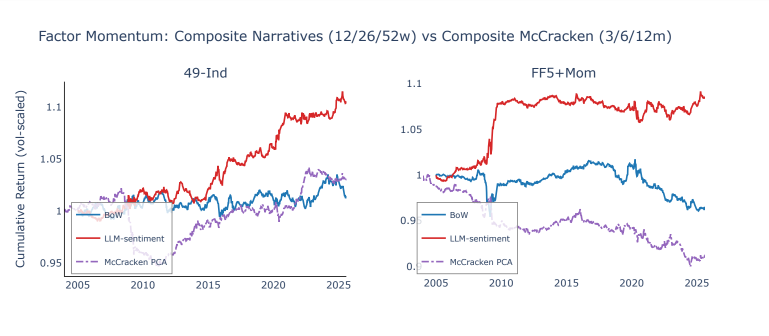
Crisis Asymmetry: GFC vs. COVID



Sharpe (sub-period)	49-Ind	FF5+Mom
GFC (slow burn)		
McCracken PCA	**0.82**	**1.43**
LLM CW	near 0 (L749)	
COVID (fast shock)		
LLM CW	1.8 to 5.6 (L751)	
McCracken PCA	0.48	−0.58

Bounded LLM scores $\Rightarrow$ $\Delta\theta \approx 0$ when sentiment stays at floor (protracted crisis). Fast shock $\Rightarrow$ text adapts within days while monthly FRED-MD lags. Sub-period SRs are noisy point estimates (GFC $\approx$78w, COVID $\approx$13w).

Narrative Momentum Equity Curves



LLM momentum: SR = 0.49 (49-Ind, $t = 2.37$), 0.39 (FF5+Mom, $t = 1.70$).

BoW and McCracken momentum: flat or negative.

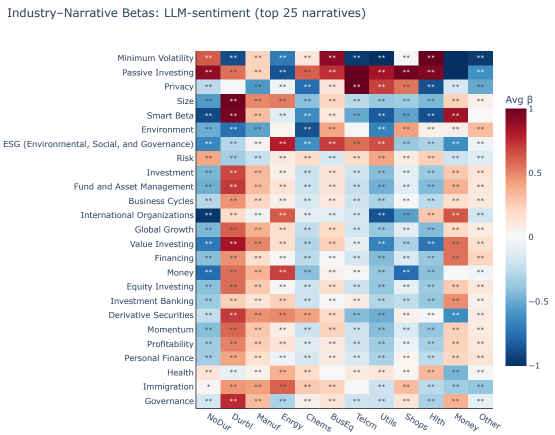
J=5, K=1, composite $\mathcal{H} = \{12, 26, 52\}$w, vol-scaled 1% ann. At implementable $K \geq 5$, signal attenuates sharply (see appendix).

Narrative Momentum, Full-Sample Summary

Model	Universe	Ann.Ret	Vol	Sharpe	NW-t
BoW	49-Ind	0.000	0.007	0.07	0.34
BoW	FF5+Mom	−0.001	0.007	−0.17	−0.69
LLM sentiment	49-Ind	0.004	0.008	0.49	2.37
LLM sentiment	FF5+Mom	0.003	0.007	0.39	1.70
McCracken PCA	49-Ind	0.001	0.008	0.15	0.70
McCracken PCA	FF5+Mom	−0.003	0.007	−0.43	−1.95

Weekly 1% vol-normalised spread returns, J=5, K=1, composite $\mathcal{H} = \{12, 26, 52\}$w. t-stats Newey–West, 8 lags. Source of truth: paper Table 3; values hand-inlined here for slide readability. Only **LLM sentiment** generates positive momentum; McCracken FF5+Mom negative.

Backup: Beta Heatmap, LLM Sentiment



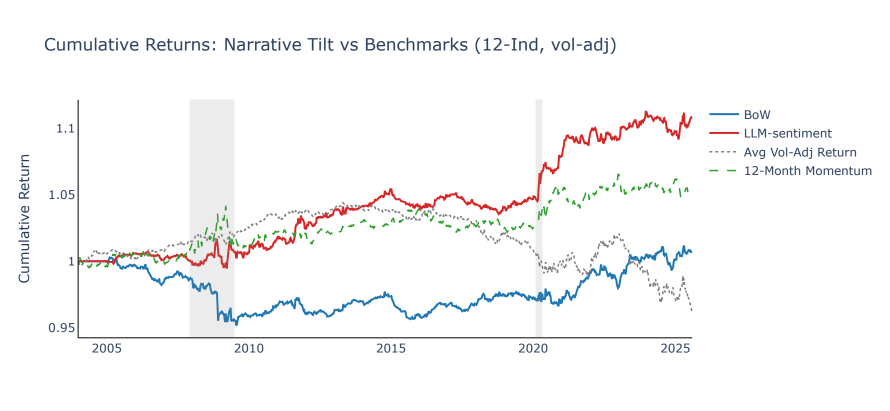
Backup: Why Would This Work?

Proposed channel: Investor inattention to gradual narrative shifts.

- Slowly evolving narratives update beliefs gradually
  (Da et al., 2014): gradual information diffusion $\Rightarrow$ underreaction
- LLM captures directional tone shifts that BoW volume misses
  (Flynn and Sastry, 2024): macro narratives affect real outcomes
- Signal exploits the gap between narrative shift and price adjustment

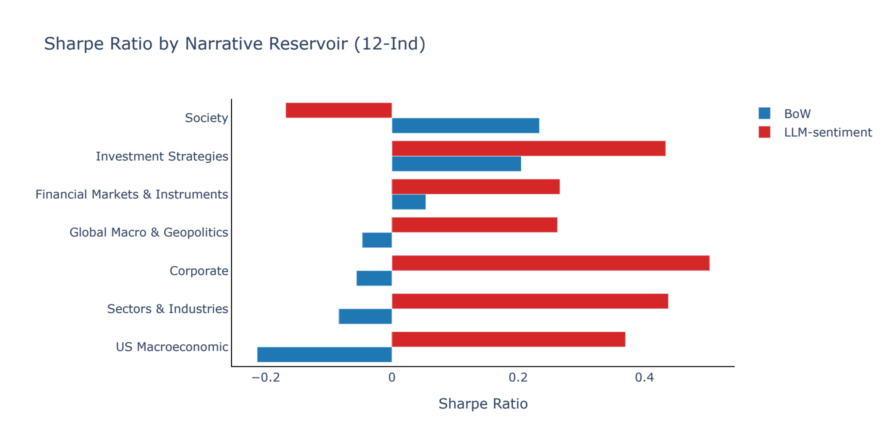
Caveat: proposed mechanism, not tested. The paper documents an *association*; the causal channel remains a hypothesis for future work.

Backup: BoW vs. LLM Direct Comparison



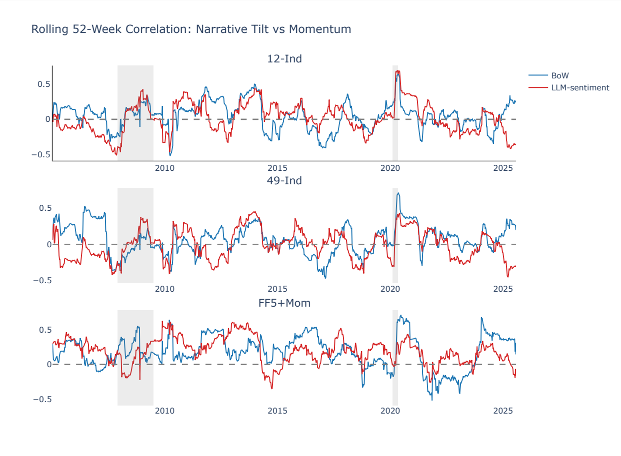
BoW attention carries negligible predictive power for cross-sectional returns.

Backup: Reservoir-Level Sharpe Comparison



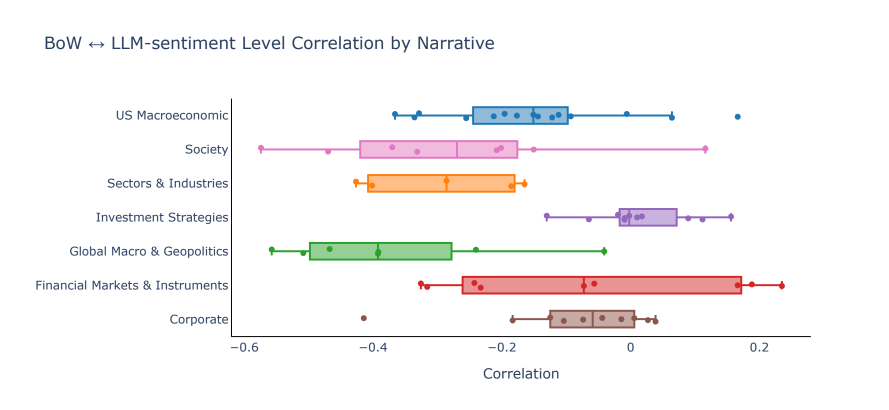
CW Sharpe ratios by narrative reservoir. No single reservoir dominates.

Backup: Rolling Correlation, CW vs. Momentum



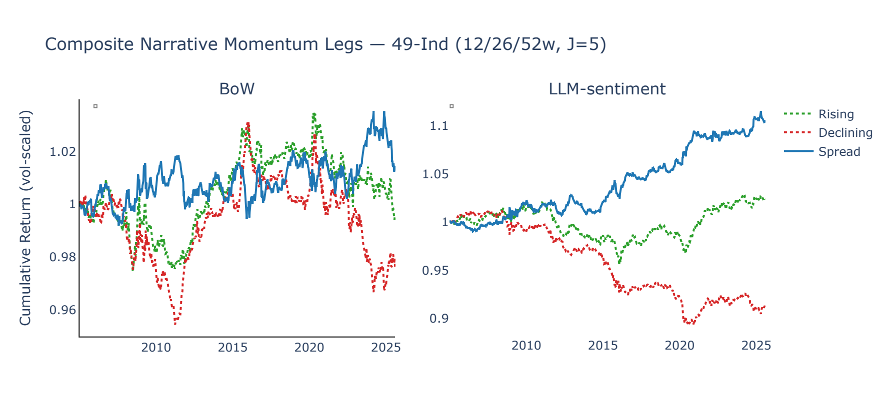
Time-varying correlation between CW and momentum strategies. Persistently low.

Backup: Cross-Model Correlation Strip



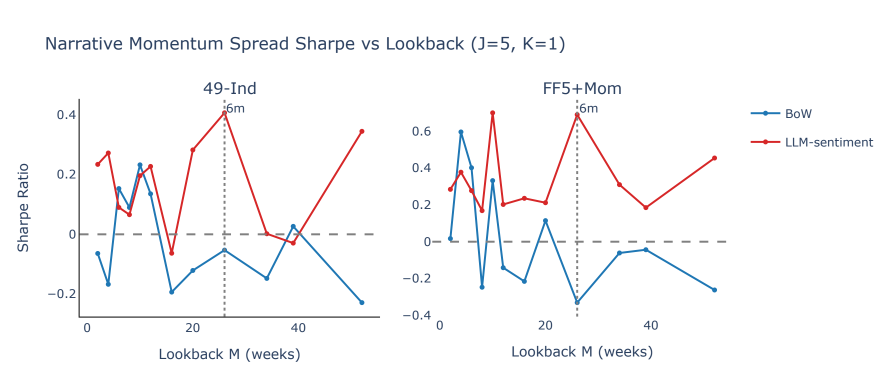
BoW–LLM change correlations by reservoir. Negative across all 7 reservoirs.

Backup: Rising vs. Declining Legs



Rising leg drives returns, identifying improving narratives is the return source.

Backup: Lookback Horizons



LLM positive across most lookbacks; BoW flat everywhere.

Composite (12/26/52 w) mitigates selection bias.

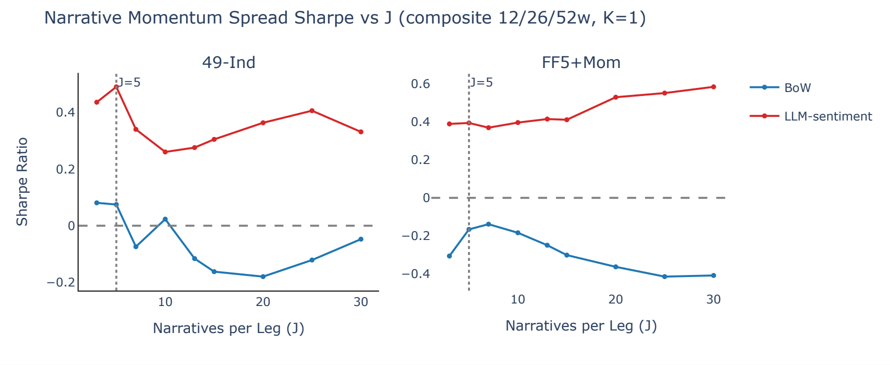
Backup: LLM provenance and prompt sensitivity

Proprietary LLM (vendor and prompt details under NDA). Prompt sensitivity untested in this paper. Beta stability is assumed across the full expanding window. The bounded-score assumption matters for crisis regimes: when sentiment saturates at the floor, $\Delta\theta \to 0$ and the signal attenuates.

Backup: Full Literature Context

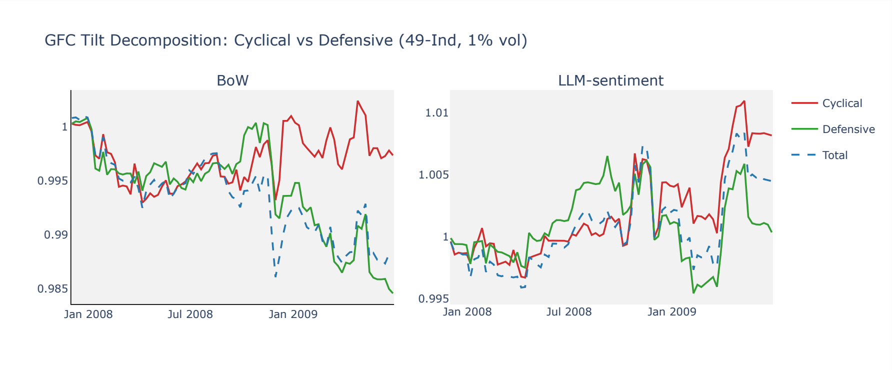
- Textual analysis in finance: Gentzkow et al. (2019); Tetlock (2007); Loughran and McDonald (2011)
- Narrative economics: Shiller (2017); Bhargava et al. (2023); Lee et al. (2024)
- Topic models & news: Bybee et al. (2024); Flynn and Sastry (2024)
- Policy uncertainty: Baker et al. (2016)
- Gradual information & momentum: Da et al. (2014)

Backup: NarrMom Spread Sharpe vs. J



Spread Sharpe as a function of J (assets per leg). $K=1$, $M=26$w. Signal is robust to moderate J but sensitive to large J.

Backup: GFC Cyclical vs. Defensive Cumulative Returns



Cumulative returns, cyclical vs. defensive industries during the GFC. Illustrates the slow-burn regime that macro PCA signals capture better than LLM sentiment.