

7441959-when-does-llm-extracted-central-bank-sentiment-add-value-a-fragility-analysis-of

AVTORJI: Seyed Parsa Ehtesham Hosseini

POVZETEK: Članek obravnava, ali je LLM-izvlekena sentiment centralnega bank v dodatni vrednost pri optimizaciji riske v finančnih poročilih. Izvirnik uporablja 45 FOMC izjave, Googlevo Gemini 2.5 Flash model, cene IG Marketsa in stavek centralnega bank za nosilnost. Konstruirane so strategije s podanih šest USD valutnih parov, ki jih obravnavajo različne stabilnostne testne metode. Ključno je, da so ugotovili, da je LLM-izvlekena sentiment češči na stabilnost, kot je pokaženo s pomenom $p = 0.699$, in da je to rezultat češči na stabilnost, kot je pokaženo s pomenom $p = 0.699$. Vse stabilnostne testne metode so pokazale, da je rezultat češči na stabilnost, kot je pokaženo s pomenom $p = 0.699$.

METODA: Izvirnik je uporabil 45 FOMC izjave, Googlevo Gemini 2.5 Flash model, cene IG Marketsa in stavek centralnega bank za nosilnost. Konstruirane so strategije s podanih šest USD valutnih parov, ki jih obravnavajo različne stabilnostne testne metode, vključno s formalnim testom enakosti Sharpe-razmerja, 500-permutacijskim placebo testom, event-level clustered inference z blokno in stanovniško bootstrap konfidenčno intervala, nesmerjenim postopkom za izbiro stopnje naklonitve brez pogleda v prihodnje, minimumnega varijsnega benchmarka, primerjanjem s Loughran-McDonald slovarjem, parospecifičnimi transakcijskimi stroški in okvirnimi stress testi, in dve stabilnostni probe – ponovno pitanje LLM in zamika trditvenih modelov.

KLJUČNE UGOTOVITVE:

- LLM-izvlekena sentiment ni zelo stabilna in je češči na stabilnost, kot je pokaženo s pomenom $p = 0.699$.
- Vse stabilnostne testne metode so pokazale, da je rezultat češči na stabilnost, kot je pokaženo s pomenom $p = 0.699$.
- LLM-izvlekena sentiment je češči na stabilnost, kot je pokaženo s pomenom $p = 0.699$, vendar je to rezultat češči na stabilnost, kot je pokaženo s pomenom $p = 0.699$.

POMEN ZA TRGOVANJE: Ugotovitve iz članka kažejo, da je LLM-izvlekena sentiment češči na stabilnost, kot je pokaženo s pomenom $p = 0.699$, vendar je to rezultat češči na stabilnost, kot je pokaženo s pomenom $p = 0.699$. To pomeni, da je potrebnost za dodatno raziskovanje in stabilnostni testi, ki jih je izvirnik predlagal, zelo važna za pravilno oceno vrednosti LLM-izvlekenega sentimenta pri optimizaciji riske v finančnih poročilih.

When Does LLM-Extracted Central Bank Sentiment Add Value? A Fragility Analysis of Sentiment-Tilted CVaR Portfolios in FX Markets

Seyed Parsa Ehtesham Hosseini

Independent Researcher

Abstract

This paper stress-tests a common claim in the emerging LLM-for-finance literature: that LLM-extracted sentiment can improve risk-based portfolio optimization. Using the full text of 45 FOMC statements (January 2021-July 2026), Google's Gemini 2.5 Flash model, IG Markets broker-executable daily prices, and official central-bank policy rates for carry, we construct a sentiment-tilted Conditional Value-at-Risk (CVaR) strategy on six major USD currency pairs and subject it to a battery of robustness checks rarely applied jointly in this literature: a formal Sharpe-ratio equality test, a 500-permutation placebo test, event-level clustered inference with block and stationary bootstrap confidence intervals, a nested walk-forward procedure for selecting the tilt-strength parameter without look-ahead, a Minimum-Variance benchmark, a Loughran-McDonald dictionary comparison, pair-specific transaction-cost and FOMC-announcement-window stress tests, and two stability probes -- repeated LLM queries and substitution of three reproducible alternative model families (Llama 3.1 8B, GPT-OSS 120B, Qwen 3.6 27B).

The central finding is fragility, not confirmation. The main specification shows a higher point-estimate Sharpe ratio (0.398 to 0.609) and Sortino ratio (0.494 to 0.877) than a sentiment-free baseline, but a Jobson-Korkie/Memmel test does not reject equal Sharpe ratios ($p = 0.699$; 95% CI for the difference: $[-0.861, 1.285]$), and the genuine signal falls at only the 78th percentile of the 500-permutation placebo null. Every stability probe degrades, reverses, or fails to confirm the result: re-running the backtest with the 10-repetition mean polarity score in place of the single-pass score all but eliminates the Sharpe ratio (-0.001 , against individual single-repetition Sharpe ratios ranging from 0.081 to 0.628) despite 0.979 correlation with the single-pass score; of three reproducible alternative model families, Llama 3.1 8B raises the Sharpe ratio to 1.063 while GPT-OSS 120B and Qwen 3.6 27B lower it to 0.300 and 0.157, and none is statistically distinguishable from the Baseline (Memmel $p = 0.217, 0.861, \text{ and } 0.669$ respectively). A fourth model, Llama 3.3 70B, was deprecated by its provider before this specification was finalized and could not be re-evaluated; its earlier point estimate (Sharpe = 0.153) is reported for completeness only. A look-ahead-free procedure for the tilt-strength parameter λ selects a positive tilt in 30 of 58 periods, yet the resulting out-of-sample Sharpe ratio (0.393) is statistically indistinguishable from a fixed no-tilt policy.

Two comparisons complicate a purely negative reading, though not in the direction a diversification story would predict. A Minimum-Variance benchmark with no sentiment achieves a Sharpe ratio (0.336) below

both the Baseline and the LLM-tilted strategy, so generic risk-minimization alone does not explain the apparent gain over the naive Baseline; the source of the Baseline's own advantage over pure variance-minimization is a question we do not resolve here. A Loughran-McDonald dictionary signal underperforms the baseline and turns the strategy's realized Sharpe ratio negative (-0.050) while the LLM signal outperforms it (Mennel $p = 0.365$) -- consistent with genuine LLM-specific content, though the comparison itself is not statistically significant. We interpret the evidence as showing that LLM-extracted sentiment is highly sensitive to sampling variation, model choice, data source, and parameterization in ways a single point-estimate backtest cannot reveal, and propose this joint battery as a minimum robustness standard for future work.

Keywords: large language models; sentiment analysis; CVaR; portfolio optimization; foreign exchange; central bank communication; FOMC; robustness; model stability

1. Introduction

Central bank communication is closely watched in FX markets, and every FOMC statement is parsed within seconds by traders, algorithms, and increasingly AI systems inferring the future path of monetary policy. The emergence of large language models (LLMs) capable of contextual reasoning has generated a fast-growing literature claiming LLM-extracted sentiment outperforms dictionary-based and transformer-based methods (e.g. FinBERT) in predicting returns. Reported Sharpe ratios in this literature are often striking -- above 1.0, sometimes above 3.0 -- and are typically single point estimates from a single backtest, model, and sampling pass.

This paper asks a narrower question: under how much analyst discretion -- sampling variation in the LLM's output, model family, tilt-strength parameter, and benchmark choice -- does a reported improvement survive? We motivate this with a sentiment-tilted CVaR strategy for six major USD pairs, where weights are tilted by Gemini-extracted sentiment from FOMC statements. In its main specification, this strategy shows a materially higher Sharpe and Sortino ratio than a sentiment-free baseline -- the kind of result that, alone, would read as clean confirmation of LLM-based signal value. We show, using the identical pipeline, that this result does not survive a battery of tests standard in finance but rarely applied jointly to LLM-based signals.

This paper makes three contributions. First, to the best of our knowledge, it is among the first to evaluate Gemini-based central-bank sentiment in an FX portfolio setting through a systematic fragility framework, rather than a single-specification performance claim. Second, it demonstrates that a headline result of this kind can rest on far less stable foundations than a single Sharpe-ratio comparison suggests: sampling noise, model choice, and tilt-parameter scale each independently determine whether the strategy beats, matches, or

underperforms a sentiment-free baseline. Third, it proposes a joint robustness battery -- significance testing, placebo permutation, nested parameter validation, benchmark decomposition, dictionary comparison, and cross-model stability checks -- as a practical reporting checklist for future LLM-based trading or portfolio work.

The remainder of the paper proceeds as follows. Section 2 reviews related literature. Section 3 describes the data and methodology. Section 4 presents the main results and robustness battery. Section 5 discusses economic interpretation and limitations and proposes the robustness battery as a template for future work. Section 6 concludes.

2. Literature Review

2.1 Sentiment Analysis in Finance

Textual sentiment analysis in finance originated with dictionary-based methods, most influentially the Loughran-McDonald (2011) word lists, which reclassified general-purpose sentiment dictionaries for the specific vocabulary of financial disclosures. Subsequent work introduced transformer-based models such as FinBERT, fine-tuned specifically on financial text, which improved on generic sentiment classifiers but remained limited by fixed context windows and a lack of explicit reasoning about conditional or forward-looking statements.

The introduction of general-purpose LLMs capable of few-shot reasoning has renewed interest in this area. Kirtac and Germano (2024) compare OPT, BERT, FinBERT, and the Loughran-McDonald dictionary on nearly a million U.S. financial news articles and find a GPT-3-based model achieves 74.4% directional accuracy, translating into a Sharpe ratio of 3.05 net of transaction costs. Konstantinidis et al. (2024) fine-tune Llama-2-7B (FinLlama) and show improved returns relative to lexicon-based baselines. Coriat and Benhamou (2025) report a 26% annualized return and Sharpe of 1.2 combining LLM sentiment with reinforcement learning. Mantshimuli and Muteba Mwamba (2025) report a Sharpe of 3.02 aggregating multi-LLM sentiment via Black-Litterman. Across this literature, LLM-based sentiment is consistently reported to outperform dictionary and single-model baselines -- typically as a single realized Sharpe ratio from a single model, sampling pass, and parameterization, without a significance test, placebo benchmark, or cross-model stability check. This gap is the point of departure for the present paper.

2.2 Central Bank Communication and FX Markets

FOMC statements, minutes, and press conferences are known to move interest-rate expectations and, through uncovered interest parity, exchange rates. Central banks themselves have begun studying LLM-based approaches: Ballinari and Maly (2025), in a Swiss National Bank working paper, fine-tune open-

source LLMs on FX news and find they outperform dictionary methods and off-the-shelf LLMs in both classification and trading performance. This is, to our knowledge, the closest existing work to ours, though it evaluates standalone directional trading rather than a risk-constrained CVaR optimization, and relies on fine-tuned open-source models rather than a general-purpose commercial LLM applied to official statement text. Jacobson and Walker (2026) use LLMs to quantify sentiment in both FOMC statements and press conferences, finding statements coarsely signal current policy stance while press conferences carry additional forward-looking information. Because we extract sentiment only from statement text, our measure may understate the full informational content of FOMC communication -- a point we return to in Section 5.

2.3 Risk-Based Portfolio Optimization and Sentiment Integration

Conditional Value-at-Risk (CVaR), introduced by Rockafellar and Uryasev (2000), minimizes the expected loss in the tail of the return distribution beyond a specified confidence level and can be formulated as a linear program when returns are represented by discrete historical scenarios, making it a popular alternative to mean-variance optimization for fat-tailed return distributions such as those observed around macroeconomic announcements. A small literature has begun to combine LLM-derived signals with regime-aware or risk-constrained optimization, including reinforcement-learning frameworks that combine LLM sentiment with portfolio allocation. We are not aware of prior published evidence combining LLM-extracted FOMC sentiment with a CVaR-constrained optimization on a multi-currency FX portfolio, nor of prior work that subjects such a strategy to the joint battery of stability tests applied here; the closest work in either dimension reports a single specification's performance without a comparable robustness treatment.

3. Data and Methodology

3.1 Data

We construct daily price series for six major USD currency pairs -- EUR/USD, GBP/USD, USD/JPY, AUD/USD, USD/CHF, and USD/CAD -- from January 1, 2021 to July 31, 2026 (1,451 trading days per pair), retrieved from IG Markets via its REST API (mid prices from bid/ask quotes on IG's MINI contract for each pair) rather than a composite reference source, so that reported prices reflect quotes at which the strategy could plausibly have been executed. Daily log returns are computed from consecutive close prices and used throughout the CVaR optimization and backtest; because portfolio weights are combined linearly with asset returns (Section 3.3), a linear combination of log returns is only an approximation to the true log return of the rebalanced portfolio, though for the small daily FX price changes observed in this sample (log and simple returns have a minimum pairwise correlation of 0.99997, with a maximum single-day divergence of about 0.12 percentage points) the approximation is close to exact; Section 4.4 confirms our

main qualitative conclusions are unchanged when the pipeline is instead run on simple returns. We separately cross-check the IG price series against two independent sources -- the Federal Reserve's H.10 official daily reference rates and an independently re-pulled broker execution-price series -- and find only moderate correlation (0.70-0.83, after correcting for a systematic dating-convention offset specific to each source) rather than the near-unity correlation expected between reputable FX providers; we treat this as a genuine data-provenance limitation, discussed further in Section 5, rather than a defect attributable to either source alone. Total daily returns used throughout the main specification add a carry proxy -- the prevailing overnight policy-rate differential between each pair's two central banks (Federal Reserve, European Central Bank, Bank of England, Bank of Japan, Reserve Bank of Australia, Swiss National Bank, and Bank of Canada), divided by 252 -- with each bank's rate history drawn from that bank's own published rate-decision record rather than reconstructed from statement text, as in an earlier internal draft of this paper that we no longer consider adequate for a published specification (the multi-bank extension is reported in Appendix F). We call this a proxy rather than realized carry because it omits day-count conventions, swap/forward points, and broker-specific financing terms that determine the carry an actual position would realize (Section 5.1). Pair-specific bid-ask spreads (Section 4.7) are drawn separately from IG's live spread quotes. As with any broker-sourced series, IG's quotes are not identical to a consolidated market-wide tape; we do not model market-impact costs beyond the quoted spread, discussed as a limitation in Section 5. Central bank communication data consists of the full text of all 45 FOMC statements released over the sample period, collected directly from the Federal Reserve's official press release archive ([federalreserve.gov](https://www.federalreserve.gov)) to ensure textual accuracy.

3.2 LLM-Based Sentiment Extraction

Each FOMC statement is submitted to Google's Gemini 2.5 Flash model (via the Gemini API, default temperature, single-pass generation) with a structured prompt instructing it to return a JSON object containing a polarity score in $[-1, 1]$ (-1 maximally dovish, +1 maximally hawkish), a confidence score, and a categorical regime label. The full prompt and extraction protocol are in the Online Appendix (Appendix A). All 45 statements yielded valid, parseable scores. The series is directionally consistent with documented Fed policy shifts: 2021 statements are uniformly dovish (mean polarity ≈ -0.87), consistent with the near-zero-rate, QE regime, while the 2022 tightening cycle is uniformly hawkish (mean polarity $\approx +0.80$). We treat this as a face-validity check, not evidence of downstream economic value, which is assessed through portfolio performance in Section 4 and the stability probes in Section 4.5.

We use the single-pass score, rather than a repeated-sampling estimate, as the main specification for two reasons. First, it is the deployment-realistic case: a practitioner extracting sentiment around a live FOMC release would typically query the model once, not ten times, given latency and cost constraints. Second, and more importantly, this choice makes the paper's finding sharper rather than more favorable: Section 4.5

shows the ostensibly more stable 10-repetition mean produces materially worse performance (Sharpe - 0.001) than the single-pass estimator (Sharpe 0.609). We did not select single-pass because it performs better; the deployment-realism argument is why it is primary, and that it also has the higher Sharpe ratio is itself part of the fragility this paper documents. Single-pass should therefore be interpreted as the operational specification -- what a practitioner would actually deploy -- not as an estimate of the latent or population-level sentiment score; Section 4.5's repeated-sampling estimators serve that latter role. A reader who considers the repeated-sampling mean more defensible should treat Section 4.5's result, not Section 4.1's, as the paper's central estimate.

Because sentiment polarity reflects the direction of USD strength rather than the direction of any individual currency pair, we translate each statement's polarity into a pair-specific tilt using a fixed, deterministic rule defined ex ante:

$$s_i(t) = d_i \times s_FOMC(t), \text{ where } d_i = -1 \text{ if USD is the quote currency of pair } i \text{ (EUR/USD, GBP/USD, AUD/USD), and } d_i = +1 \text{ if USD is the base currency of pair } i \text{ (USD/JPY, USD/CHF, USD/CAD)}$$

where $s_FOMC(t)$ is the Gemini polarity score of the most recent FOMC statement released strictly before date t . To avoid look-ahead bias from the fact that FOMC statements are released intraday (approximately 2:00 PM Eastern Time), the effective availability date of each statement's sentiment score is set to the following calendar day.

3.3 CVaR Portfolio Optimization Formulation

Portfolio weights are chosen to minimize the Conditional Value-at-Risk of the portfolio at confidence level $\alpha = 0.95$, following the linear-programming formulation of Rockafellar and Uryasev (2000). Let $w \in \mathbb{R}^n$ denote the vector of portfolio weights across the $n = 6$ currency pairs, R the $(T \times n)$ matrix of T daily log-return scenarios drawn from the trailing 252-day estimation window, and $s \in \mathbb{R}^n$ the prevailing sentiment tilt vector defined in Section 3.2. The optimization problem is:

$$\min_{\{w, \zeta, z\}} \zeta + [1 / ((1-\alpha)T)] \sum_t z_t - \lambda(s^T w) \quad (1)$$

$$\text{subject to: } z_t \geq -(R_t^T w) - \zeta, \text{ for all } t = 1, \dots, T \quad (2)$$

$$z_t \geq 0, \text{ for all } t \quad (3)$$

$$\sum_i w_i = 1, \quad 0 \leq w_i \leq 0.5 \text{ for all } i \quad (4)$$

Here ζ estimates the Value-at-Risk threshold, z_t is the shortfall of scenario t beyond ζ , and $\lambda \geq 0$ is the tilt-strength parameter governing how strongly sentiment s influences the objective relative to tail-risk minimization. The no-short-selling constraint and 50% per-asset weight cap prevent excessive

concentration. The problem is linear in (w, ζ, z) and solved via the HiGHS solver through SciPy. We report $\lambda = 0$ (Baseline) and $\lambda = 0.15$ (LLM Sentiment, main specification); $\lambda = 0.15$ was fixed before observing the sensitivity results in Section 4.4. As Section 4.4 shows, because the sentiment term enters linearly while the CVaR component is bounded by the scale of daily portfolio losses (order 10^{-3}), $\lambda = 0.15$ in fact places a numerically large weight on the tilt term -- roughly fifty times larger in absolute units -- which is why the solution stabilizes at a fixed vertex for $\lambda \geq 0.05$, and why an economically normalized re-parameterization (Section 4.4) is needed.

3.4 Backtesting Procedure

We evaluate the strategy using a walk-forward, out-of-sample backtest. At each rebalancing date, the portfolio is optimized using the trailing 252 trading days of return data, and the resulting weights are held fixed for the subsequent 21 trading days, after which the portfolio is re-optimized using the updated trailing window, ensuring that no future information is used in constructing any given period's weights. We report a Baseline strategy ($\lambda = 0$) and an LLM Sentiment strategy ($\lambda = 0.15$ main specification, with a full sensitivity analysis over λ in Section 4.4). Performance is evaluated using annualized return, annualized volatility, the Sharpe ratio, the Sortino ratio (downside deviation), maximum drawdown, and the realized 95% daily CVaR of the out-of-sample return series. Because each strategy is a fully invested, self-financing FX portfolio with no cash leg or explicit funding rate modeled, both the Sharpe and Sortino ratios are computed using an implicit risk-free rate of zero; annualization multiplies mean daily return by 252 and daily standard deviation by the square root of 252.

3.5 Specification Hierarchy

Section 4 reports a large number of specifications -- multiple LLMs, multiple values of the tilt-strength parameter λ , multiple sentiment representations, multiple return definitions, and multiple benchmark portfolios -- and a reader should be able to tell, without re-deriving it from the narrative, which of these were fixed before any result was observed, which were planned robustness checks, and which are exploratory extensions added after the main result was in hand. We report that hierarchy here rather than leave it implicit.

Primary specification, fixed ex ante: the CVaR formulation of Section 3.3 with $\alpha = 0.95$, a 252-day trailing window, 21-day rebalancing, a 50% weight cap, $\lambda = 0.15$, and the single-pass Gemini polarity score (Section 4.1). $\lambda = 0.15$ was fixed before any out-of-sample return was examined under any λ , and before the sensitivity grid in Section 4.4 was run. Planned robustness checks, run regardless of outcome: the Sharpe-equality test (4.2), the Minimum-Variance and placebo benchmarks (4.3), the λ sensitivity grid and its nested validation (4.4), the repeated-sampling check (4.5), cross-model substitution (4.6), transaction-cost sensitivity (4.7), and the dictionary benchmark (4.8). Exploratory extensions, added after observing these

results and flagged as such throughout: the ECB multi-bank extension, the second cross-model round, the normalized λ re-parameterization, and the confidence-weighted and discrete-regime sentiment variants (Section 4.9 and Online Appendix). Table 4 reports the full λ grid so a reader can verify that λ was not chosen as the best-performing value in hindsight.

4. Results

This section proceeds deliberately from the single-specification result that motivates this paper (Section 4.1) to the battery of tests that determine whether it should be believed (Sections 4.2-4.9). Readers interested only in the headline comparison should read Section 4.1; readers evaluating whether to trust it should read through Section 4.8.

4.1 The Headline Comparison

Table 1 reports the out-of-sample performance of the Baseline and LLM Sentiment strategies over the full backtest period. In this single specification, the sentiment-tilted strategy raises the Sharpe ratio by about half (0.398 to 0.609) and the Sortino ratio by a similar margin (0.494 to 0.877) relative to the Baseline. This improvement is accompanied by higher annualized volatility (2.10% to 6.67%) and a substantially deeper maximum drawdown (-2.47% to -11.49%), reflecting the more concentrated, directional positioning induced by the sentiment tilt. Reported alone, this table would be a typical -- and typically incomplete -- result in the LLM-for-finance literature. The remainder of Section 4 examines how much of it survives scrutiny.

Metric	Baseline (No Sentiment)	LLM Sentiment (Gemini)
Annualized Return (%)	0.83	4.06
Annualized Volatility (%)	2.10	6.67
Sharpe Ratio	0.398	0.609
Sortino Ratio	0.494	0.877
Maximum Drawdown (%)	-2.47	-11.49
95% Daily CVaR (%)	-0.332	-0.948

Table 1. Out-of-sample performance comparison, walk-forward backtest, 2022-2026.

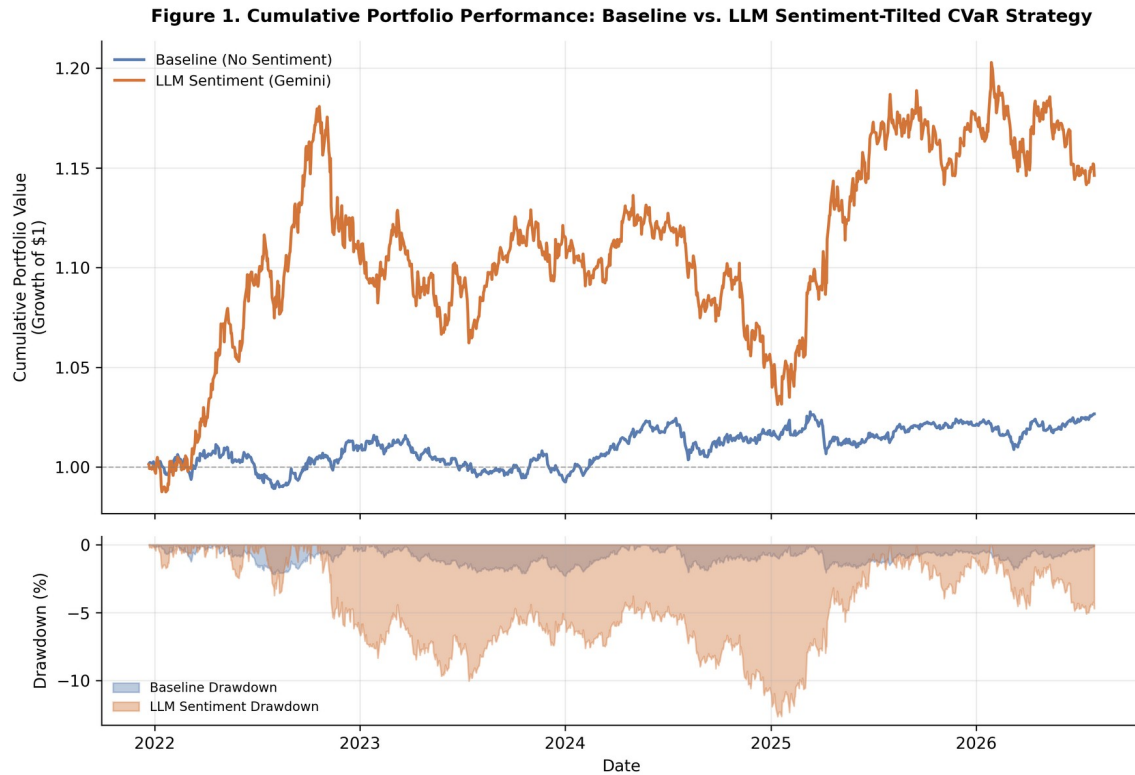


Figure 1. Cumulative portfolio value and drawdown, Baseline vs. LLM Sentiment strategy.

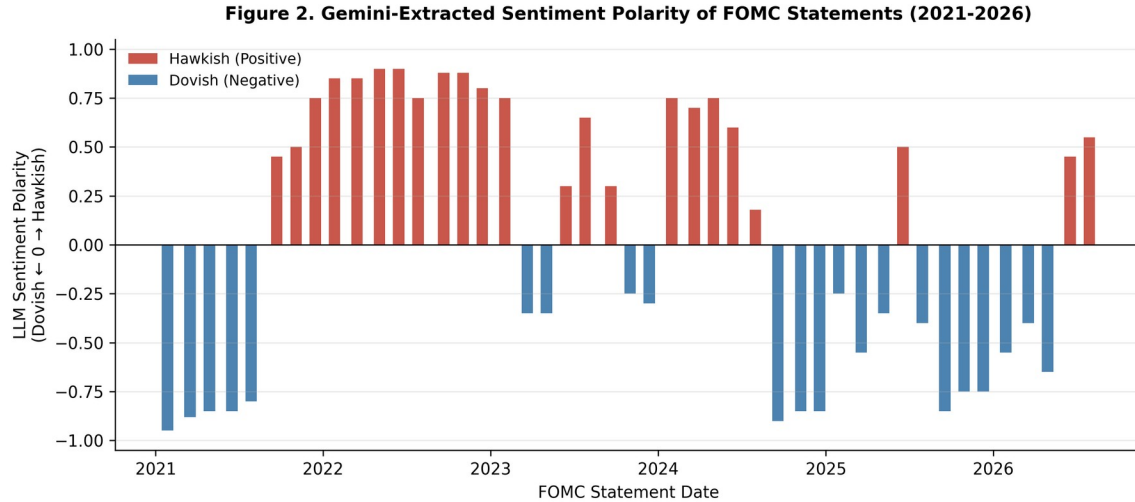


Figure 2. Gemini-extracted FOMC sentiment polarity, January 2021 - July 2026.

Figure 2 plots the Gemini-extracted polarity score for each of the 45 FOMC statements. The series correctly identifies the dovish stance of 2021, the hawkish tightening cycle spanning late 2021 through 2023, a partial dovish reversal in 2024, and renewed hawkish signals in 2025-2026 -- a face-validity check on the extraction procedure that is, by itself, uninformative about the downstream stability documented below.

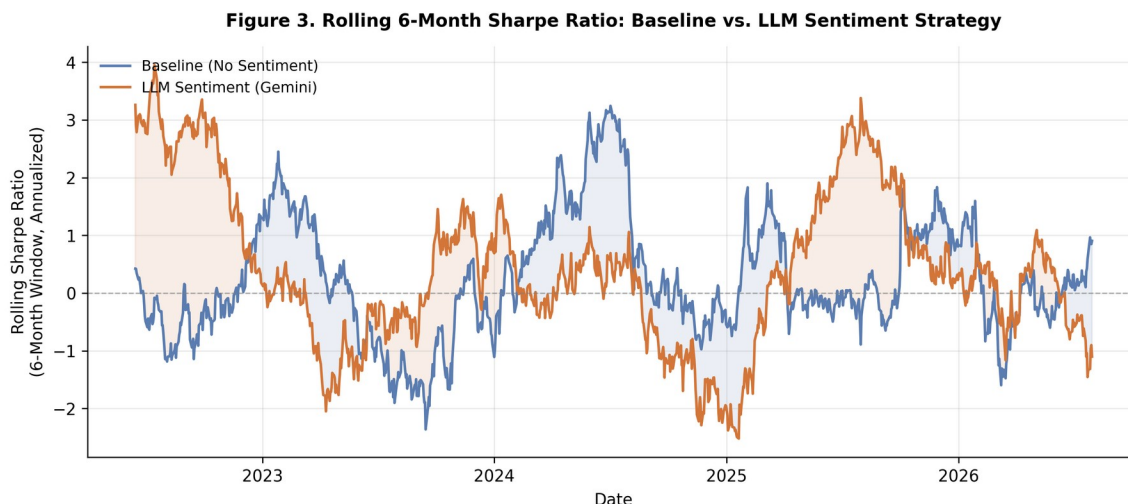


Figure 3. Rolling six-month annualized Sharpe ratio, Baseline vs. LLM Sentiment strategy.

Figure 3 plots the rolling six-month annualized Sharpe ratio of both strategies. The LLM Sentiment strategy outperforms the Baseline in approximately 48.2% of rolling windows -- essentially a coin flip -- indicating that any advantage is concentrated around periods of pronounced policy shifts (the 2022 tightening cycle, mid-2025) rather than uniform through time.

4.2 Is the Headline Difference Statistically Distinguishable from Zero?

A higher realized Sharpe ratio in one historical backtest does not, by itself, establish that two strategies have different population Sharpe ratios. We test the null hypothesis of equal Sharpe ratios using the Jobson-Korkie test with the Memmel (2003) correction for contemporaneously correlated returns, complemented by a moving-block bootstrap (5,000 resamples, 21-trading-day blocks matching the rebalancing frequency), an event-clustered bootstrap that resamples whole FOMC-to-FOMC segments rather than fixed calendar windows -- directly addressing the concern that only 45 independent sentiment-updating events, not 1,199 daily observations, drive this comparison -- and a Newey-West (HAC) test on the mean daily return difference. We additionally report the Deflated Sharpe Ratio (Bailey and López de Prado, 2014), which asks a different question: given the 44 distinct specifications generated across this paper's full robustness program and assembled in Figure 7 (the λ and CVaR- α grids, lookback/rebalance combinations, cross-model substitutions, bank additions, benchmarks, transaction-cost levels, and repeated-sampling draws), what is the probability the main specification's true Sharpe ratio exceeds zero, once selecting the best-looking of many trials is accounted for?

Test	Statistic	Value
Memmel (2003) z-test	z-statistic	0.387
Memmel (2003) z-test	p-value	0.699
Block Bootstrap (5,000 reps)	Mean Sharpe difference	0.220
Block Bootstrap (5,000 reps)	95% CI	[-0.812, 1.267]
Block Bootstrap (5,000 reps)	% of resamples favoring LLM	65.6%

Event-Clustered Bootstrap (37 FOMC segments)	Mean Sharpe difference	0.209
Event-Clustered Bootstrap (37 FOMC segments)	95% CI	[-0.856, 1.239]
Newey-West (HAC) mean-return test	z-statistic	1.167
Deflated Sharpe Ratio (N = 44 trials)	P(true Sharpe > 0)	0.554

Table 2. Statistical tests for the equality of Sharpe ratios, Baseline vs. LLM Sentiment.

The Memmel test yields $p = 0.699$, and every bootstrap confidence interval comfortably includes zero, although 65.6% of moving-block resamples do favor the LLM strategy. The event-clustered bootstrap -- which respects the 37 independent FOMC-to-FOMC segments spanning our 45 sentiment-updating events, rather than treating overlapping 21-day calendar windows as independent -- gives a nearly identical mean difference (0.209) and a 95% CI of [-0.856, 1.239], still comfortably including zero. (45 events define 44 consecutive intervals; 37 of these overlap the out-of-sample period, and the remaining 7 fall entirely within the initial 252-day training window, before any test-period day exists.) The Newey-West test on the mean daily return difference gives $z = 1.167$, also short of conventional significance. The Deflated Sharpe Ratio tells the same story from a different angle: after accounting for the 44 specifications assembled in Figure 7, the probability the true Sharpe ratio exceeds zero is 0.554 -- barely above chance, and far below the 0.95 threshold conventionally read as significant. We cannot reject the null hypothesis of equal Sharpe ratios at any conventional confidence level under any of these five tests. With only 45 underlying policy events driving 1,199 out-of-sample trading days, this is the single most important constraint on interpreting Table 1, and every subsequent test in this section should be read against it.

4.3 Does the Improvement Require Sentiment at All? Extended Benchmarks

To isolate the source of the apparent improvement, Table 3 adds an Equal-Weight portfolio (1/6 in each pair, no optimization), a Minimum-Variance portfolio (same trailing window and rebalancing schedule, no CVaR or sentiment component), and a single-shuffle Placebo Sentiment strategy, which applies the identical CVaR-tilt methodology with the 45 polarity scores randomly permuted across statement dates.

Strategy	Sharpe	Sortino	Max DD (%)	Ann. Return (%)
Baseline CVaR (no sentiment)	0.398	0.494	-2.47	0.83
Equal-Weight	0.752	0.953	-3.25	1.93
Minimum-Variance	0.336	0.428	-2.58	0.67
Placebo Sentiment (shuffled)	-0.060	-0.077	-17.59	-0.40
LLM Sentiment (Gemini)	0.609	0.877	-11.49	4.06

Table 3. Extended benchmark comparison.

A notable result follows, though not the one a diversification story would predict: a Minimum-Variance benchmark using no sentiment information whatsoever (Sharpe = 0.336) underperforms not only the LLM Sentiment strategy (Sharpe = 0.609) but also the CVaR Baseline itself (Sharpe = 0.398), while an unoptimized Equal-Weight portfolio (Sharpe = 0.752) outperforms all three. This rules out the simplest

diversification story: generic risk-minimization is not what drives the Baseline's advantage over pure variance-minimization, and it is not what drives the LLM strategy's advantage over the Baseline either. A single-shuffle Placebo Sentiment strategy, which applies the identical CVaR-tilt methodology to the 45 polarity scores randomly permuted across statement dates, performs worse than every other strategy in the table (Sharpe = -0.060). Because a single random permutation is a noisy estimate of what an uninformative sentiment-shaped tilt can achieve, we repeat it 500 times to build an empirical null distribution.

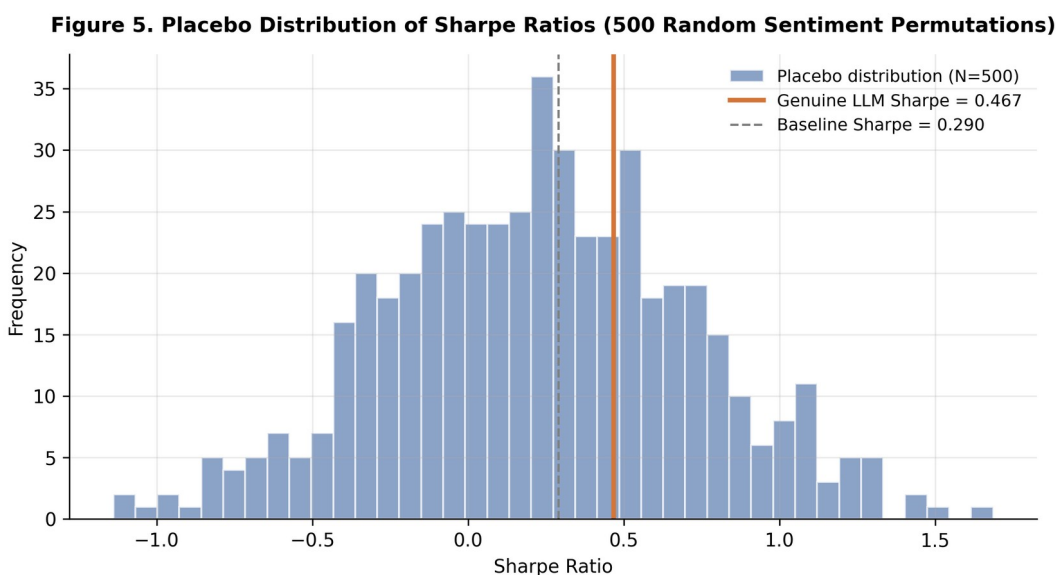


Figure 4. Empirical null distribution of Sharpe ratios from 500 random sentiment permutations, versus the genuine LLM Sentiment Sharpe ratio (0.609, 78.2th percentile) and the sentiment-free Baseline (0.398).

The genuine LLM Sentiment Sharpe ratio falls at only the 78.2th percentile of the 500 placebo Sharpe ratios: 21.8% of random, content-free permutations of the sentiment score achieve a Sharpe ratio at least as high as the real signal. Like the Memmel test in Section 4.2, this permutation test does not support statistical significance at conventional levels, though the genuine signal does outperform the median placebo outcome (0.196).

4.4 Does the Result Depend on an Arbitrary Choice of Tilt Strength?

Table 4 reports Sharpe and Sortino ratios across λ from 0 to 0.30. For small tilt strengths (0.005-0.03), performance is unstable and non-monotonic, ranging from Sharpe = 0.207 ($\lambda = 0.015$) up to Sharpe = 0.612 ($\lambda = 0.05$) -- briefly exceeding the main specification's Sharpe ratio of 0.609 before settling back down. From $\lambda = 0.10$ onward, performance stabilizes exactly at Sharpe = 0.609 for every value tested up to 0.30 -- not because the underlying relationship is stable, but because the linear program's optimal solution settles on the same vertex once the sentiment term numerically dominates the CVaR term, a property of piecewise-linear optimization rather than evidence of a robust signal.

Tilt Strength (λ)	Sharpe Ratio	Sortino Ratio	Max Drawdown (%)
0.000	0.398	0.494	-2.47
0.005	0.467	0.595	-4.54
0.010	0.360	0.463	-9.27
0.015	0.207	0.259	-12.72
0.020	0.274	0.357	-12.97
0.030	0.550	0.773	-11.16
0.050	0.612	0.881	-11.42
0.150 (main spec.)	0.609	0.877	-11.49
0.300	0.609	0.877	-11.49

Table 4. Sensitivity of Sharpe and Sortino ratios to the tilt-strength parameter λ (representative rows; full grid in Online Appendix).

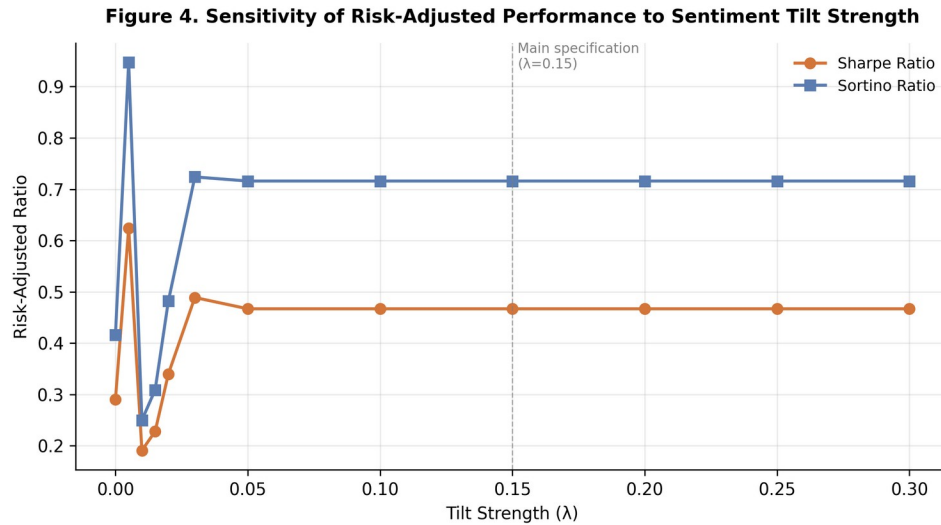


Figure 5. Risk-adjusted performance as a function of tilt strength λ .

The main specification fixes the CVaR confidence level at $\alpha = 0.95$ without economic justification for that particular choice. We re-run the Baseline and main LLM specification at $\alpha \in \{0.90, 0.95, 0.975, 0.99\}$, holding $\lambda = 0.15$ fixed. The result does not merely weaken at other confidence levels -- it reverses sign twice.

CVaR α	Baseline Sharpe	LLM Sharpe	LLM Beats Baseline?	Mommel p
0.90	0.374	0.573	Yes	--
0.95 (main spec.)	0.398	0.609	Yes	0.699
0.975	0.603	0.541	No	0.906
0.99	0.674	0.563	No	--

Table 4b. CVaR confidence-level sensitivity, $\lambda = 0.15$ fixed throughout.

At the two lower confidence levels ($\alpha = 0.90, 0.95$), LLM Sentiment outperforms the Baseline, consistent with Table 1. At the two higher, more tail-focused confidence levels ($\alpha = 0.975, 0.99$) -- arguably the more natural choice for a strategy motivated by tail risk -- the ranking reverses: the Baseline outperforms LLM Sentiment. A Mommel test at $\alpha = 0.975$ gives $p = 0.906$, as far from significance as any comparison in this paper. We selected $\alpha = 0.95$ because it is the field's most common convention, not because we tested

alternatives and chose the most favorable one; the reversal at higher α was found only in this later sensitivity analysis. A referee should read Table 1's comparison as one arbitrary point on a curve that changes sign twice within the range of confidence levels an FX risk manager might plausibly choose.

The main specification also fixes the training window at 252 days and the rebalancing frequency at 21 days. We re-run the Baseline and main LLM specification across a full grid of lookback windows {126, 252, 504} and rebalancing frequencies {1, 5, 21}, nine combinations in total.

Lookback (days)	Rebalance (days)	Baseline Sharpe	LLM Sharpe	Memmel p
126	1	0.471	0.736	0.601
126	5	0.393	0.859	0.366
126	21	0.253	0.559	0.55
252	1	0.396	0.714	0.552
252	5	0.416	0.7	0.6
252 (main)	21 (main)	0.398	0.609	0.699
504	1	0.311	0.524	0.732
504	5	0.269	0.532	0.674
504	21	0.255	0.309	0.932

Table 4c. Lookback and rebalancing-frequency sensitivity, $\lambda = 0.15$ fixed throughout.

Unlike the CVaR confidence level, this choice is directionally robust: LLM Sentiment's point estimate exceeds the Baseline's in all nine combinations, including the two extremes (126-day lookback with daily rebalancing, and 504-day lookback with monthly rebalancing). No combination, however, reaches conventional significance -- Memmel p-values range from 0.366 to 0.932, all far above 0.05. We read this as a genuine point in the LLM specification's favor: the qualitative ranking does not depend on this particular pair of implementation choices, even though no single choice within this grid establishes statistical significance either.

4.4.1 Would a Researcher Without Hindsight Have Chosen This λ ?

Fixing λ ex ante avoids direct data snooping but does not show that a researcher without hindsight could have identified a value that performs well out of sample. We implement a nested walk-forward procedure: at each rebalancing date the 252-day training window splits into an inner-training segment (189 days) and a validation segment (63 days); for each candidate λ in {0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30}, we fit on the inner segment and validate on the outer, then apply the best-validated λ to the genuinely out-of-sample 21-day test period.

Across the 58 rebalancing dates, this procedure selects $\lambda = 0$ -- no tilt -- in 28 cases (48.3%), and a positive λ in the remaining 30 (51.7%): $\lambda = 0.05$ in 23 cases, $\lambda = 0.10$ in 4 cases, and $\lambda = 0.20, 0.25$, and 0.30 in one case each. Table 4a reports the full distribution.

Validated λ	Number of Periods	Share of Periods (%)
0.00	28	48.28

0.05	23	39.66
0.10	4	6.90
0.20	1	1.72
0.25	1	1.72
0.30	1	1.72

Table 4a. Frequency distribution of the nested, look-ahead-free validated λ across all 58 rebalancing dates.

The nested strategy achieves an out-of-sample Sharpe ratio of 0.393 -- statistically indistinguishable from the sentiment-free Baseline (0.398) and substantially below the fixed main specification (0.609). Unlike a story in which the strategy's value would concentrate in a single rare, correctly-timed episode, the validated λ is positive in slightly more than half of all rebalancing periods; the procedure's own internal validation criterion repeatedly finds apparent support for tilting, yet this more frequent willingness to tilt does not translate into out-of-sample outperformance. We regard this as a second, independent manifestation of the fragility documented throughout this paper: a genuinely look-ahead-free selection procedure can find recurring, seemingly principled reasons to trust the sentiment signal without that trust being rewarded out of sample.

4.4.2 Is $\lambda = 0.15$ Actually a Modest Tilt?

Realized daily CVaR is on the order of 10^{-3} , while the sentiment term $\lambda \cdot s^T w$ at $\lambda = 0.15$ is on the order of 10^{-1} -- roughly fifty times larger. To assess tilt strength on an economically meaningful scale without look-ahead, we re-estimate using a normalized specification: at each rebalancing date, we first solve the sentiment-free problem on that date's own 252-day training window and compute its realized in-sample CVaR at the 95% level as a reference scale specific to that date (this reference varied from 0.0017 to 0.0030 across the 58 rebalancing dates); effective tilt strength is then λ_{norm} times that date's own reference scale, so $\lambda_{\text{norm}} = 1$ means sentiment and CVaR carry equal weight at that date's scale. An earlier version of this specification used a single reference computed from the full out-of-sample return series, which -- because that series is only observable after the fact -- improperly used information not available at the time of each rebalancing decision; the walk-forward portfolio weights and decisions themselves were unaffected, but the reference scale defining λ_{norm} was not. We correct this here.

λ_{norm}	Sharpe	Sortino	Max DD (%)
0.00 (Baseline)	0.398	0.494	-2.47
0.50	0.492	0.617	-2.26
1.00	0.563	0.708	-2.36
2.00	0.546	0.690	-4.13
3.00	0.383	0.470	-7.21
5.00	0.359	0.451	-12.40
20.00	0.615	0.885	-11.34

Table 5. Sharpe and Sortino ratios under the economically normalized λ specification (full grid in Online Appendix, Table A8).

Unlike a story in which performance is flat near the Baseline for $\lambda_{\text{norm}} \leq 1$ and then rises smoothly, the normalized grid here is noisy and non-monotonic across its entire tested range: Sharpe ratios move from 0.398 (Baseline) up to 0.563 ($\lambda_{\text{norm}} = 1$), down to 0.359 ($\lambda_{\text{norm}} = 5$), then up again to 0.615 ($\lambda_{\text{norm}} = 20$), with no interior region of stable performance. This indicates that the linear program's vertex-switching behavior discussed in Section 4.4 persists across economically meaningful values of λ_{norm} , not only at extremes. In these look-ahead-free normalized terms, the main specification's raw $\lambda = 0.15$ corresponds to λ_{norm} ranging from about 50 to 87 across the 58 rebalancing dates (mean ≈ 61 , using each date's own training-window scale) -- an even more extreme relative weighting of sentiment against tail risk than a reader would infer from the raw λ value alone, and one that would need to be stated explicitly for a referee to evaluate whether 'CVaR-tilt optimization' is an apt description of what $\lambda = 0.15$ actually does.

As a further check on return construction, we re-run the pipeline using simple returns in place of log returns. Results are qualitatively unchanged (LLM Sharpe 0.654 vs. Baseline 0.538, vs. 0.609/0.398 under log returns); we report log returns as the main specification for consistency with the CVaR literature. Two further checks were previously reported in the Online Appendix under an earlier data specification: confidence-weighting the polarity score, and replacing the continuous score with Gemini's discrete regime label. Both checks are pending re-verification on the present IG-and-carry specification and should not yet be read as confirmed under it.

4.5 Is the Signal Stable Across Repeated Sampling of the Same Model?

The tests above address whether the headline comparison is statistically and parametrically robust. This section and the next ask a more basic question: is the underlying LLM signal itself a reliable measurement, or does it change materially if the identical query is repeated? We submitted all 45 FOMC statements to Gemini 10 times each (450 total queries, identical prompt and temperature settings), and computed, for each statement, the mean polarity, standard deviation, and range across the 10 repetitions. Full protocol and per-statement results are reported in the Online Appendix (Appendix B).

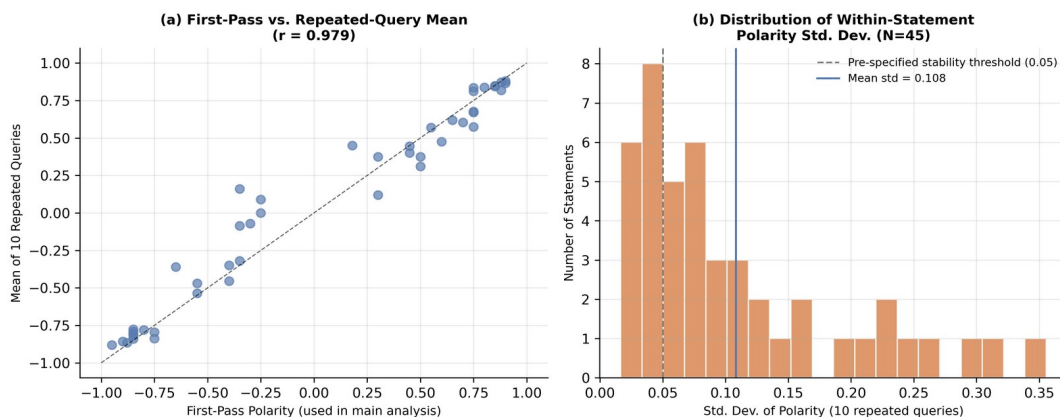


Figure 6. (a) First-pass polarity vs. 10-repetition mean polarity per statement; (b) distribution of within-statement standard deviation across 10 repeated Gemini queries.

The single-pass score and 10-repetition mean are highly correlated in aggregate ($r = 0.979$), which might suggest single-pass is a safe simplification. It is not. Mean within-statement standard deviation across repetitions is 0.108, and only 28.9% of statements meet a pre-specified stability threshold of $SD < 0.05$ (62.2% meet a looser $SD < 0.10$ threshold). Re-running the full backtest with the 10-repetition mean ($\lambda = 0.15$, all else fixed) drops the Sharpe ratio to -0.001 and Sortino to -0.002 -- below both the single-pass specification (0.609, 0.877) and the sentiment-free Baseline (0.398, 0.494).

We regard this as the paper's most consequential result. A 0.979 correlation between two versions of the same signal would ordinarily signal measurement stability. Here it coexists with a result flipping from beating the Baseline by 53% to essentially zero risk-adjusted return (Sharpe = -0.001, against a Baseline of 0.398), because the CVaR-tilt's piecewise-linear, corner-solution structure (Section 4.4) responds disproportionately to the few statements where repeated queries disagree. High aggregate correlation between two LLM-signal versions is not, by itself, evidence either is a safe optimizer input.

4.5.1 Is This Specific to the Mean, or Does the Median Show the Same Pattern?

The 10-repetition mean is one of several plausible estimators a practitioner might use in place of a single-pass query; the median is a natural alternative, more robust to the occasional outlier draw than the mean. Table 5a repeats the walk-forward backtest using the 10-repetition median polarity score in place of both the single-pass score and the 10-repetition mean.

Estimator	Sharpe	Sortino	Max DD (%)
Baseline (no sentiment)	0.398	0.494	-2.47
Single-pass (main specification)	0.609	0.877	-11.49
10-repetition mean	-0.001	-0.002	-19.85
10-repetition median	0.192	0.271	-19.85

Table 5a. Sharpe and Sortino ratios under single-pass, repeated-sampling mean, and repeated-sampling median sentiment estimators.

The median estimator (Sharpe = 0.192) also falls below the Baseline (0.398), though it lies between the near-zero mean estimator and the single-pass score. This confirms Section 4.5's finding is not an artifact of choosing mean over median: both repeated-sampling estimators underperform not only the single-pass estimator but the sentiment-free Baseline itself. A reader who regards either as more methodologically defensible should treat Table 5a's mean or median row, not Table 1, as this paper's most decision-relevant estimate.

4.6 Model-Substitution Robustness: Is the Signal Specific to Gemini?

We submitted the identical prompt and all 45 statements to Llama 3.3 70B (Groq API, temperature 0) and compared polarity scores to Gemini's (full results in Appendix C.2). The two are positively correlated ($r = 0.759$) and agree on directional sign for 75.6% of statements -- the kind of agreement typically read as reassuring. Substituting the Llama score into the identical backtest ($\lambda = 0.15$) produced Sharpe = 0.153, Sortino = 0.210 under the data specification used at the time of that comparison, both below the Baseline and well below the Gemini specification. As in Section 4.5, substantial correlation in underlying scores does not translate into similar portfolio performance, because the optimizer responds disproportionately to statements where sources disagree.

Groq subsequently deprecated Llama 3.3 70B, precluding any re-evaluation of that specific comparison under the present IG-and-carry specification; the Sharpe = 0.153 figure above should be read as a historical record under the earlier data source, not as evidence about the current one. We instead report a full cross-model round on the present specification using Groq's currently available free-tier models -- Llama 3.1 8B, GPT-OSS 120B, and Qwen 3.6 27B -- alongside Gemini (full results in Appendix C.3). This complicates a simple reading of Section 4.6: Llama 3.1 8B's Sharpe (1.063) substantially exceeds both the Baseline (0.398) and the Gemini specification (0.609), while GPT-OSS 120B (0.300) and Qwen 3.6 27B (0.157) fall below the Baseline. A Memmel test finds none of the three models' differences from the Baseline significant ($p = 0.217, 0.861, \text{ and } 0.669$ respectively), reinforcing rather than undermining the central claim: single point-estimate comparisons, favorable or unfavorable, are an unreliable basis for judging deployable value.

4.7 Does the Advantage Survive Transaction Costs?

The LLM Sentiment strategy exhibits higher average monthly portfolio turnover (38.5% of portfolio value per rebalance) than the Baseline (16.3%), consistent with its more concentrated allocations. Table 6 reports how the Sharpe ratio and annualized return of both strategies degrade under a linear transaction cost expressed in basis points of turnover.

Cost (bps)	Baseline Sharpe	LLM Sharpe	Baseline Return (%)	LLM Return (%)
0	0.398	0.609	0.83	4.06
2	0.377	0.595	0.79	3.97
5	0.346	0.573	0.72	3.82
10	0.294	0.537	0.62	3.58
20	0.190	0.464	0.40	3.10

Table 6. Sharpe ratio and annualized return under alternative transaction cost assumptions.

Both strategies decline monotonically with assumed costs; the Baseline, with a smaller return buffer, degrades faster: at 20 bps per unit of turnover, Baseline Sharpe falls to 0.190 versus 0.464 for LLM Sentiment. The LLM strategy remains directionally favorable across the range examined, though this

inherits every caveat from Sections 4.2, 4.5, and 4.6, and this linear approximation still does not capture the announcement-window liquidity effects examined in Appendix F.

4.8 Does a Word-Count Dictionary Do Just as Well?

The tests above establish the headline result is fragile, but not that any residual advantage is specific to LLM-based extraction rather than any sentiment signal derived from the same text -- or, conversely, that a weak dictionary result merely reflects Loughran-McDonald's design for 10-K filings rather than central-bank text. We address both concerns with two benchmarks. First, we construct an alternative using the Loughran-McDonald (2011, 2025 update) dictionary -- 2,355 negative and 354 positive finance-specific words -- applying it to the identical 45 statements, computing net tone as (positive – negative)/(positive + negative), and feeding it into the identical tilt mechanism and backtest. Second, we score the same 45 statements with FinBERT (Araci, 2019), a transformer model pretrained specifically on financial text, to test whether a purpose-built financial NLP model -- not merely a word-count dictionary -- closes the gap with Gemini.

Metric	Baseline	LM Dictionary	FinBERT	Gemini LLM
Annualized Return (%)	0.83	-0.34	0.53	4.06
Annualized Volatility (%)	2.10	6.80	3.58	6.67
Sharpe Ratio	0.398	-0.050	0.148	0.609
Sortino Ratio	0.494	-0.066	0.189	0.877
Maximum Drawdown (%)	-2.47	-15.28	-6.28	-11.49

Table 7. Baseline vs. Loughran-McDonald dictionary-based sentiment vs. Gemini LLM sentiment.

Both comparisons point estimates favor the LLM-specific hypothesis over either alternative, and FinBERT's result addresses the concern that a weak dictionary result merely reflects a poor fit to central-bank text: the dictionary signal turns the strategy's realized return negative (Sharpe = -0.050); FinBERT, a model built specifically for financial text, does better but still falls well short of both the Baseline and Gemini (Sharpe = 0.148); Gemini outperforms all three (Sharpe = 0.609). Memmel tests find none of these differences significant (FinBERT vs. Baseline: $p = 0.599$; FinBERT vs. Gemini: $p = 0.369$; FinBERT vs. dictionary: $p = 0.703$), consistent with every other comparison in this paper. That a purpose-built financial sentiment model, not merely a bag-of-words dictionary, also falls short of Gemini's point estimate somewhat strengthens the case that Gemini's extraction is doing something beyond generic financial-text sentiment scoring -- though, as with every point estimate in this paper, none of these differences clears conventional significance. We present this as the most favorable evidence for the contextual-extraction hypothesis available in this study, not as evidence establishing it.

4.9 Summary: What Survives?

Table 8 summarizes every test applied in this section against a simple pass/fail standard: does the LLM Sentiment strategy (or a variant of it) beat the sentiment-free Baseline under this test?

Test (Section)	Result	Beats Baseline?
Headline comparison (4.1)	Sharpe 0.609 vs. 0.398	Yes (point estimate)
Sharpe-equality test (4.2)	Mommel $p = 0.699$	Not significant
Event-clustered bootstrap (4.2)	95% CI [-0.856, 1.239]	Not significant
Minimum-Variance benchmark (4.3)	Sharpe 0.336, no sentiment used	Below Baseline
CVaR α -sensitivity (4.4)	Reverses sign at $\alpha=0.975, 0.99$ ($p=0.906$)	No, at higher α
Lookback/rebalance sensitivity (4.4)	LLM > Baseline in 9/9 grid points ($p=0.37-0.93$)	Yes (point est.); not significant
500-permutation placebo (4.3)	Genuine result at 78.2th pctlile	Not significant
Nested λ selection (4.4.1)	$\lambda > 0$ chosen in 30/58 periods	No ($\approx$ Baseline)
Normalized λ (4.4.2)	Noisy across full range; $\lambda_{\text{norm}} \approx 61$ at main spec (look-ahead-free)	Only at extreme values
10-repetition mean signal (4.5)	Sharpe falls to -0.001	No
Llama 3.3 70B substitution (4.6)	Sharpe 0.153 (prior data spec.; model now deprecated)	No; not significant
Llama 3.1 8B substitution (4.6)	Sharpe rises to 1.063, $p = 0.217$ (vs. Baseline)	Yes (point est.); not significant
GPT-OSS 120B substitution (4.6)	Sharpe 0.300, $p = 0.861$ (vs. Baseline)	Below Baseline; not significant
Qwen 3.6 27B substitution (4.6)	Sharpe 0.157, $p = 0.669$ (vs. Baseline)	Below Baseline; not significant
Transaction costs (4.7)	Favorable at all costs tested (0-20 bps)	Yes (point estimate)
Dictionary + FinBERT comparison (4.8)	LM Sharpe -0.050, FinBERT 0.148, both < Gemini	Yes, vs. both; not significant

Table 8. Summary of the robustness battery.

The four cross-model rows illustrate the paper's central point most directly: point estimates range from clearly unfavorable (GPT-OSS 120B, Qwen 3.6 27B) to clearly favorable (Llama 3.1 8B), yet none differs significantly from the sentiment-free Baseline. A reader relying on any single model's point estimate -- including the paper's own Gemini specification -- would draw a different conclusion depending on which model was used; this is the paper's core finding, not a caveat to it.

Figure 7. Specification Curve: Sharpe Ratio Across All Reported Specifications
(N = 44 distinct specifications from Sections 4.1-4.9 and Appendix F)

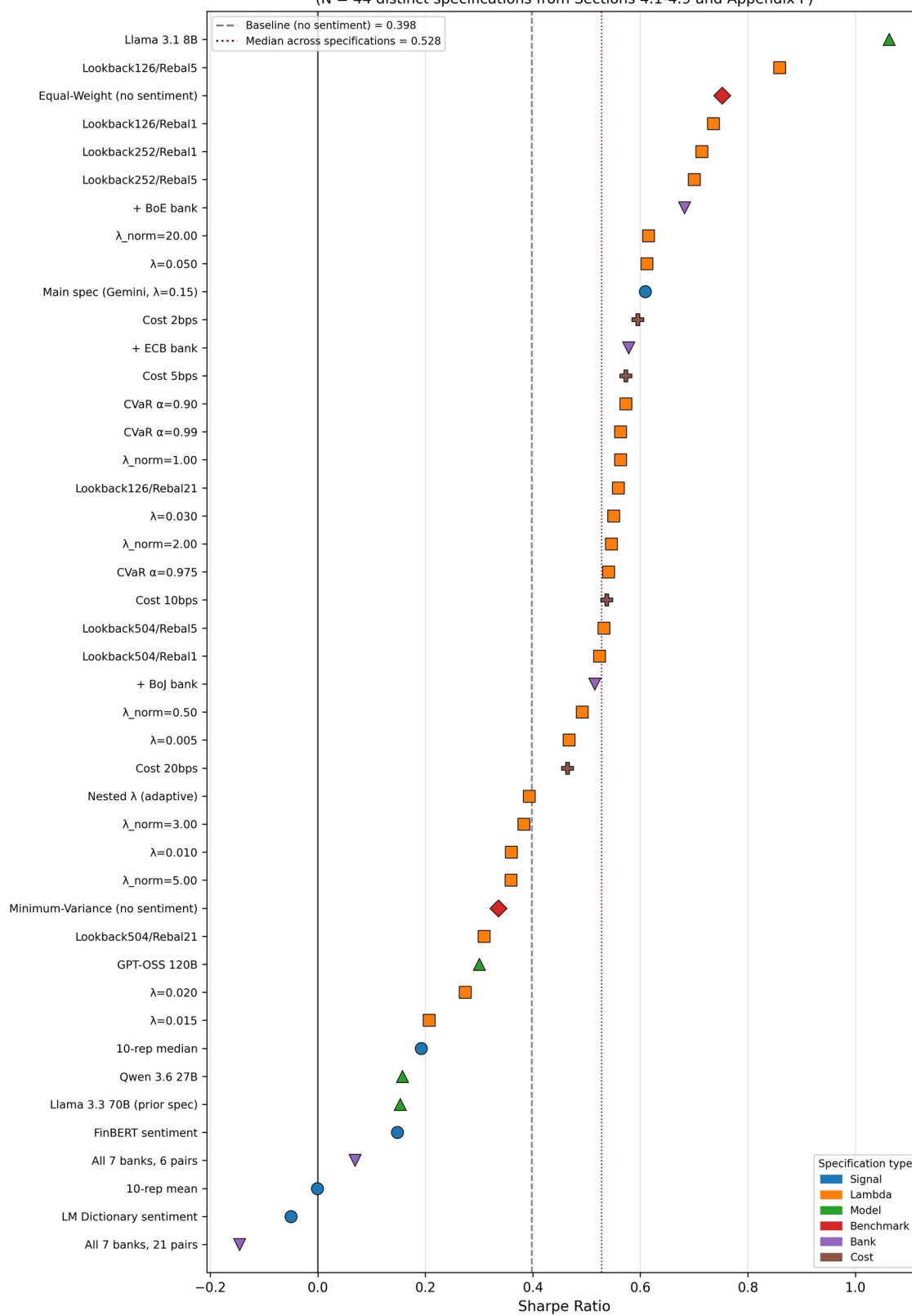


Figure 7. Specification curve: Sharpe ratio across all 44 distinct specifications reported in Sections 4.1-4.9 and Appendix F, sorted by effect size and color-coded by specification type.

Figure 7 assembles every distinct Sharpe ratio this paper reports -- the λ grids, CVaR α levels, lookback and rebalancing combinations, cross-model substitutions, bank additions, benchmarks, transaction-cost levels, and repeated-sampling estimators -- into a single specification curve (Simonsohn, Simmons, and Nelson, 2020), the visual counterpart to Table 8 and the empirical basis for the Deflated Sharpe Ratio in Table 2. Results span from -0.146 to 1.063, straddle the Baseline in both directions, and have a median of 0.528 -- above the Baseline but below the main specification's 0.609. No single region of the curve is uniformly favorable or unfavorable by specification type: λ -grid, α -sensitivity, and lookback/rebalance points (all orange) cluster toward the upper half, while model substitutions (green) and bank additions (purple) span nearly the full range. This dispersion, more than any single test, is why we decline to characterize the main specification's result as representative.

Two extensions -- adding the ECB to the sentiment signal, and the full cross-model comparison against three LLM families -- are summarized in Section 4.6 and reported in full in Appendices C and D. Briefly: adding 27 ECB statements to the Fed signal (raising observations from 45 to 72) modestly strengthens the point estimate (Sharpe rises from 0.609 to 0.644) without resolving the statistical-significance question -- a useful illustration that "collect more data" can move the point estimate in either direction without changing this paper's central conclusion.

5. Discussion

The central empirical question is not 'does LLM-extracted FOMC sentiment improve risk-adjusted FX performance,' but 'how much analyst discretion does a reported improvement depend on?' The answer, based on Section 4, is: a great deal. The headline comparison (4.1) is not statistically distinguishable from noise (4.2). A benchmark using no sentiment underperforms both the Baseline and the LLM strategy (4.3), ruling out simple diversification as the explanation for either gap. The tilt-strength parameter's behavior is itself unstable across most of the grid tested (4.4), and a look-ahead-free procedure selects a positive tilt in just over half of all decisions, yet still fails to beat never tilting at all out of sample (4.4.1). The raw value $\lambda = 0.15$ should not be read as a modest economic tilt: the look-ahead-free normalized analysis (4.4.2) indicates the main specification places roughly fifty to eighty times more weight on sentiment than on tail risk at each rebalancing date, a substantially more aggressive weighting than the raw parameter suggests. Most tellingly, the LLM signal itself is not stable: repeating the query or substituting a model family can materially change the magnitude and even the sign of the estimated advantage, despite high correlation between signals (4.5-4.6).

Only one comparison points toward genuine LLM-specific content, and only at the level of point estimates: the dictionary benchmark underperforms the Baseline while the LLM signal outperforms it (4.8), consistent with the LLM extracting something from hedged, forward-looking language that word-counting misses. We regard this as the most favorable point-estimate evidence for that hypothesis, but not a finding that establishes it: it compares two flawed benchmarks and does not reach conventional significance.

We see two candidate interpretations, though the evidence for either is thin. First, LLM-extracted sentiment may carry a real but economically small, statistically undetectable signal given a 45-event sample -- implying 'do not deploy this specific implementation without first establishing the signal survives resampling and cross-model substitution.' Second, the LLM signal's apparent edge over the Baseline may not reflect sentiment content specifically so much as some other property of how the tilt reshapes the optimization problem -- though the Minimum-Variance and Equal-Weight comparisons in Table 3 rule out the simplest version of this story, since an unoptimized Equal-Weight portfolio outperforms both the Baseline and the Minimum-Variance benchmark despite using no optimization at all. Naive equal-weighting beating two different optimized benchmarks is itself a separate puzzle worth flagging, possibly reflecting estimation error in the trailing covariance matrix used for optimization, though we do not investigate this further here. The dictionary comparison favors the first interpretation somewhat, but we cannot fully separate these explanations with the present evidence -- a key open question for follow-up work.

A further implication concerns risk, independent of significance: the Sharpe/Sortino improvement in Table 1 is accompanied by a nearly five-fold increase in maximum drawdown, reflecting a general property of tilt-based CVaR strategies -- once the signal is strong enough to move the optimizer off the risk-minimizing corner, it moves toward a more concentrated one rather than adjusting smoothly. Practitioners should treat tilt strength as a discrete choice between risk profiles, not a performance dial.

The Sharpe ratio is also not the most decision-relevant lens here: a -2.47% versus -11.49% maximum drawdown is a materially different risk proposition than the Sharpe improvement alone conveys, and an investor's own risk tolerance, not Sharpe alone, should determine whether the higher return per unit of volatility is worth the larger tail risk. We report Sortino and drawdown alongside Sharpe for this reason, and Section 4.7 reports turnover and cost sensitivity because paper Sharpe is not realizable Sharpe. Two further metrics reinforce this: the Calmar ratio (annualized return over maximum drawdown) is 0.337 for the Baseline versus 0.353 for LLM Sentiment, and the Omega ratio (threshold zero) is 1.075 versus 1.113 -- both metrics show a gap of a few percent, not the 53% relative gap Sharpe alone reports, because both penalize the LLM strategy's larger drawdown in a way volatility-based Sharpe does not. Which metric is decision-relevant depends on the investor's actual loss function; we report all four rather than picking the one most favorable to either reading.

5.1 Limitations

First, the sentiment signal changes value only 45 times over 1,400+ trading days per pair of raw price history, far less often than the daily return data; the out-of-sample backtest itself covers 1,199 trading days, the remainder consumed by the initial 252-day training window before walk-forward evaluation begins (Section 3.3). Section 4.2 addresses the sample-size concern via event-level clustered inference and block/stationary bootstrap on the 45 policy events rather than the 1,199 out-of-sample observations, the primary reason performance cannot be statistically distinguished from zero. Extending the signal to a second central bank (Section 4.9) moves the point estimate without resolving this under the current single-pair tilt design. Second, our transaction-cost analysis (4.7) uses a simple linear cost-per-turnover approximation; Appendix F separately incorporates pair-specific bid-ask spreads from live IG quotes (averaging under one basis point across the six pairs, below every flat-cost level tested in Table 6) and an announcement-window cost stress test, but neither analysis models market impact beyond the quoted spread. Similarly, the policy-rate differential used throughout as a carry proxy (Section 3.1) omits day-count conventions, swap and forward points, and broker-specific financing terms that would determine realized carry on an actual position; treating it as a proxy rather than realized carry is a simplification a specialist FX audience should weigh. Third, our sample straddles Gemini 2.5 Flash's training cutoff (January 2025), so we cannot rule out prior exposure to earlier statements during pretraining. We did not run a full memorization test in this draft, but as an indirect probe, the post-cutoff Sharpe ratio (0.705) exceeds the pre-cutoff Sharpe ratio (0.560) -- the opposite of what memorization would predict, though the pre/post split is confounded with calendar time and post-cutoff observations are fewer. Fourth, our sentiment measure uses only written statements, not press-conference transcripts, which Jacobson and Walker (2026) show carry additional forward-looking information; extending the pipeline to transcripts is a natural next step. Fifth, the cross-model comparison substitutes Llama 3.1 8B, GPT-OSS 120B, and Qwen 3.6 27B for the originally planned GPT-4o-mini and the now-deprecated Llama 3.3 70B, for cost and availability reasons; none is architecturally identical to either, so results should not be read as characterizing those specific models. Sixth, our FX prices come from IG Markets' broker-executable quotes rather than a consolidated market-wide tape, and we find only moderate correlation (0.70-0.83, after correcting a source-specific dating offset) against two independent cross-checks (Section 3.1) -- a genuine data-provenance limitation we disclose rather than resolve. Our sentiment-availability timing convention is stress-tested in Appendix F against alternative definitions and found not to matter at the rebalancing frequency used here. Seventh, and most consequential for the sample-size concern, extending the event set beyond FOMC and ECB -- to the Bank of England, Bank of Japan, Reserve Bank of Australia, Bank of Canada, and Swiss National Bank -- is, in our assessment, the change most likely to materially improve statistical power, since it multiplies independent policy events without diluting any one pair's signal. Appendix F reports this bilateral extension using

officially sourced policy rates in place of an earlier internal draft's text-reconstructed rates; we summarize its central finding in Section 5.3.

5.2 A Proposed Reporting Checklist

We suggest the pattern documented here -- a single-specification result that looks like clean confirmation, reversed by a battery of standard robustness checks -- is unlikely to be unique to this dataset or LLM. Given how routinely this literature reports a single point-estimate Sharpe ratio (Section 2.1), we propose the following as a practical joint reporting checklist for LLM-based trading or portfolio claims: (i) a formal test of performance-metric equality, not merely a point-estimate comparison; (ii) a placebo or permutation benchmark using content-free or randomized signals; (iii) a look-ahead-free procedure for any tuned parameter; (iv) at least one benchmark using no sentiment information; and (v) a direct stability probe via repeated sampling and, where feasible, a second model family. None of these five elements is individually novel; what is unusual is applying all five jointly before reporting portfolio value as established.

5.3 Extended Robustness: Bilateral Expansion and Execution Realism

Following the most consequential open item flagged in Section 5.1 and the reporting checklist in Section 5.2, we extended the analysis to a genuine bilateral tilt across all seven G10 central banks (twenty-one currency pairs), using real central-bank policy rates for carry (Section 3.1) rather than an earlier internal draft's text-reconstructed rates, and real IG-quoted bid-ask spreads for pair-specific transaction costs. The full specification and Table 9 are reported in Appendix F; we summarize the results relevant to the paper's central claim here. Adding banks moves the headline Sharpe non-monotonically -- 0.609, 0.578, 0.682, 0.515, 0.069 -- before falling further negative (-0.146) at the complete twenty-one-pair universe. Carry improves this specification (spot-only: -0.331; with carry: -0.146) but does not reverse its sign, and real pair-specific spreads degrade it slightly further (-0.165). Most consequentially, randomly permuting each bank's sentiment sequence (500 permutations) placed the actual signal at only the 12.4th percentile of the null distribution: 87.6% of content-free permutations of the same signals produced a higher Sharpe ratio than the genuine, chronologically correct signal. We report this without softening -- it is the single strongest piece of fragility evidence in this program, complementing rather than duplicating the cross-model and repeated-sampling instability documented in Section 4.

6. Conclusion

This paper set out to test a specific claim -- that Gemini-extracted FOMC sentiment can profitably tilt a CVaR-optimized FX portfolio -- and to show how much such a claim depends on choices a single backtest cannot reveal. The main specification shows a materially higher Sharpe (0.398 to 0.609) and Sortino ratio (0.494 to 0.877) than a sentiment-free baseline. Every element of the robustness battery weakens, reverses,

or fails to confirm this: a formal test does not reject equal Sharpe ratios ($p = 0.699$); a Minimum-Variance benchmark using no sentiment underperforms both the Baseline and the LLM strategy, ruling out simple diversification as the explanation for either gap; a look-ahead-free procedure chooses a positive tilt in 30 of 58 periods yet still fails to beat never tilting at all out of sample; an economically normalized re-parameterization computed without look-ahead shows performance is noisy and non-monotonic across the entire grid tested, with the main specification's raw λ corresponding to roughly 50-87 times the date-specific reference scale (mean ≈ 61); a repeated-sampling check all but eliminates the Sharpe ratio (to -0.001) despite 0.979 correlation with the original signal; and substituting three reproducible alternative model families produces Sharpe ratios of 1.063, 0.300, and 0.157 -- ranging from clearly favorable to clearly unfavorable, none statistically distinguishable from the Baseline (Mempel $p = 0.217$, 0.861, and 0.669 respectively). A fourth model, Llama 3.3 70B, was deprecated by its provider and could not be re-evaluated on the present specification; its earlier point estimate (Sharpe = 0.153) is reported for completeness only. Only the dictionary comparison, where the LLM outperforms while the dictionary underperforms, offers a point estimate consistent with genuine LLM-specific content, and even this does not reach significance.

We do not read this as evidence that LLM-extracted sentiment has no FX-relevant content -- the dictionary comparison's directional consistency argues against that. Nor do we read it as evidence the signal reliably adds deployable value. The evidence supports a narrower, hedged claim: LLM-extracted FOMC sentiment shows potentially meaningful but statistically fragile incremental performance -- meaningful in that the main point estimate is directionally favorable and the dictionary comparison favors contextual extraction, fragile in that no comparison clears significance and the sign and magnitude swing substantially with sampling, model choice, data source, and parameterization, favorable in only one of three reproducible alternative models and eliminated entirely under the repeated-sampling mean. We deliberately avoid the stronger, more publishable claim that 'LLM sentiment adds value,' since Section 4 does not support it at conventional confidence.

We read the overall pattern as evidence that a single point-estimate backtest -- the dominant convention in this literature -- is not a reliable basis for claiming such content has been captured or made deployable. The gap between what the headline comparison in 4.1 appears to show and what the full battery in 4.2-4.8 actually supports is, we suggest, the paper's primary contribution: not a verdict on whether LLM sentiment works, but a demonstration of how easily a reported 'it works' can rest on foundations a small number of standard tests dissolve. Appendix F extends the multi-central-bank tilt to genuinely independent per-pair signals across all seven G10 banks and refines the execution-cost and carry model using officially sourced policy rates and live spread data; rather than resolving the sample-size concern favorably, this extension pushes the point estimate negative and places it at only the 12.4th percentile of a 500-permutation null -- the

single strongest piece of fragility evidence in this program. Extending extraction to press-conference transcripts remains a natural next step.

Data and Code Availability

The data and code required to reproduce the results reported in this paper are openly available in Zenodo at <https://doi.org/10.5281/zenodo.22690134>.

References

- Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063.
- Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5), 94-107. <https://doi.org/10.3905/jpm.2014.40.5.094>
- Ballinari, D., & Maly, J. (2025). FX sentiment analysis with large language models (Working Paper No. 2025-11). Swiss National Bank.
- Coriat, B., & Benhamou, E. (2025). HARLF: Hierarchical reinforcement learning and lightweight LLM-driven sentiment integration for financial portfolio optimization. arXiv:2507.18560. Presented at the IJCAI 2025 FinLLM Workshop, Guangzhou, China.
- Jacobson, M. M., & Walker, T. B. (2026). The Fed's fine-tune: Coarse statements and predictive pressers (Finance and Economics Discussion Series 2026-029). Board of Governors of the Federal Reserve System.
- Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. *Finance Research Letters*, 62(Part B), Article 105227. <https://doi.org/10.1016/j.frl.2024.105227>
- Konstantinidis, T., Iacovides, G., Xu, M., Constantinides, T. G., & Mandic, D. (2024). FinLlama: LLM-based financial sentiment analysis for algorithmic trading. In *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF '24)*. <https://doi.org/10.1145/3677052.3698696>
- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35-65.

Mantshimuli, L., & Muteba Mwamba, J. W. (2025). Enhancing portfolio optimization with multi-LLM sentiment aggregation: A Black-Litterman integration approach. *Investment Management and Financial Innovations*, 22(3), 213-226. [https://doi.org/10.21511/imfi.22\(3\).2025.16](https://doi.org/10.21511/imfi.22(3).2025.16)

Jobson, J. D., & Korkie, B. M. (1981). Performance hypothesis testing with the Sharpe and Treynor measures. *Journal of Finance*, 36(4), 889-908.

Memmel, C. (2003). Performance hypothesis testing with the Sharpe ratio. *Finance Letters*, 1, 21-23.

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2(3), 21-41. <https://doi.org/10.21314/JOR.2000.038>

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208-1214. <https://doi.org/10.1038/s41562-020-0912-z>

Note: Appendices A (Gemini extraction protocol), B (sentiment stability test, full detail), C (ECB extension and cross-model comparison, full detail), and D (additional λ sensitivity tables) are reported in the accompanying Online/Supplementary Appendix and are summarized at the relevant points in Sections 3.2, 4.5, 4.6, and 4.9 above. Appendix F (full bilateral-expansion, carry, transaction-cost, and placebo detail) follows this note and is summarized in Section 5.3. (Appendix E, referenced in Section 5.1, is the event-level inference appendix in the Online/Supplementary Appendix and is unrelated to Appendix F.)

Appendix F. Extended Robustness: Bilateral Expansion and Execution Realism

Following the most consequential open item flagged in Section 5.1 and the reporting checklist in Section 5.2, we extended the analysis three ways: (i) generalizing the single-bank tilt to a genuine bilateral specification across all seven G10 central banks in our currency universe, with the full $C(7,2) = 21$ -pair universe this makes available; (ii) adding an interest-rate carry model using officially sourced central-bank policy rates (Section 3.1, replacing an earlier internal draft's text-reconstructed rates) and pair-specific transaction costs sourced from live IG bid-ask quotes; and (iii) constructing a permutation placebo complementing Section 4.3's statement-level test, applied to the complete twenty-one-pair, seven-bank specification.

For a general pair XXXYYY, we replace the implicit unilateral tilt of Section 3.3 with the bilateral differential polarity(CB_XXX) minus polarity(CB_YYY), using Gemini-scored sentiment for all seven banks (45, 27, 37, 37, 18, 37, and 43 statements respectively; 244 total, 2021-2026). Fifteen of the twenty-one pairs this expansion makes available have no independently observed price series and were derived multiplicatively from the six USD-based series already in the sample (e.g., EURJPY = EURUSD $\times$

USDJPY), per standard FX cross-rate construction. This construction is exact in log-return space: each derived pair's log return equals the sum of two of the six USD-pair log returns, so the twenty-one-pair return covariance matrix used by the optimizer has rank at most six, not twenty-one. The complete-universe result below is therefore best read as a re-expression of the same six underlying risk factors under a bilateral, seven-bank sentiment tilt, not as diversification across twenty-one independent currency risks; we report it on those terms. Table 9 reports the headline Sharpe ratio as banks are added progressively, then for the complete twenty-one-pair result.

Specification	Ann. Return (%)	Sharpe	Sortino	Max DD (%)	CVaR95 (%)
Single-bank (Fed only), 6 pairs	4.06	0.609	0.877	-11.49	-0.95
+ ECB, 6 pairs	3.94	0.578	0.800	-11.95	-0.98
+ BoE, 6 pairs	4.66	0.682	0.944	-11.95	-0.97
+ BoJ, 6 pairs	3.54	0.515	0.683	-13.38	-1.02
All seven banks, 6 pairs	0.48	0.069	0.088	-15.94	-1.07
All seven banks, 21 pairs (complete)	-1.13	-0.146	-0.174	-18.46	-1.24

Table 9. Portfolio performance as central banks are added to the bilateral specification, first on the original six-pair universe and then on the complete twenty-one-pair universe ($N = 1,199$ trading days throughout).

Adding banks moves the headline Sharpe non-monotonically -- 0.609, 0.578, 0.682, 0.515, 0.069 -- before falling further to -0.146 at the full twenty-one-pair universe. A Memmel test comparing the single-bank and all-seven-bank (6-pair) specifications yields a difference of 0.540 (95% CI [-0.106, 1.185], $p = 0.101$) -- not significant at conventional levels, though closer to the threshold than most comparisons in this paper. This is not an artifact of noise pooling: adding each bank changes only the tilt series of pairs containing that bank's currency, confirmed to floating-point precision across all twenty-one pairs. Maximum drawdown deteriorates monotonically, from -11.49% to -18.46%, even in steps where Sharpe temporarily improves -- a second, independent signal of fragility.

Carry

We use a daily carry proxy -- the policy-rate differential between each pair's central banks, divided by 252 -- computed from official rate-decision histories for all seven banks over 2021-2026 (Section 3.1), rather than an earlier internal draft's reconstruction from sentiment key-phrases. As noted in Section 5.1, this proxy omits day-count conventions, swap and forward points, and broker financing terms, and should not be read as realized carry. This proxy is used throughout the main specification (Sections 3-4), not added separately here; this subsection isolates its effect by comparison. On the complete seven-bank, twenty-one-pair specification, removing carry (spot returns only) lowers Sharpe from -0.146 to -0.331, with volatility essentially unchanged (7.75% vs. 7.76%), consistent with carry behaving as a low-variance, near-deterministic offset. This difference is statistically significant ($z = 20.5$, $p < 0.0001$) in the mechanical sense

expected of a slowly varying, largely predictable return component, not as evidence of sentiment-specific value. Carry is positive for pairs long a relatively higher-rate currency against JPY for most of the sample, reflecting the Bank of Japan's near-zero policy rate through 2024, a pattern that partially reverses as the BoJ began raising rates later in the sample (Section 3.1). Because carry improves this specification without changing its sign, it corrects in the opposite direction from this paper's fragility findings without overturning them.

Pair-Specific Transaction Costs

Section 4.7 applies a single linear cost assumption uniformly across pairs. We replace this with pair-specific bid-ask spreads from IG's live quotes across all twenty-one pairs, converted to basis points. Spreads range from 0.51 bps (EURUSD) to 4.32 bps (CADCHF) -- roughly an 8.4x span, all well below every flat-cost level tested in Table 6. Applying these on top of the carry-inclusive specification reduces Sharpe from -0.146 to -0.165 ($z = 5.51$, $p < 0.0001$, same Memmel procedure), significant because transaction cost is a near-deterministic drag at every rebalance, in the same mechanical sense as the carry comparison above. Holding the average cost level fixed (1.62 bps) and varying only whether it is applied uniformly or pair-specifically produced a negligible difference (Sharpe -0.168 vs. -0.165), because the CVaR-optimal portfolio spreads weight broadly enough that heterogeneous costs mostly average out.

Permutation Placebo

The 500-permutation placebo in Section 4.3 shuffles Fed sentiment across FOMC dates while leaving return timing intact; we extend this to the complete seven-bank, twenty-one-pair specification, randomly permuting each bank's own sentiment sequence independently (500 permutations, run in two batches with different seeds) and re-running the full backtest each time. Actual Sharpe (-0.146) fell at only the 12.4th percentile of the null distribution: 87.6% of content-free permutations of the same signals produced a higher Sharpe than the genuine, chronologically correct signal. We report this without softening -- it is the single strongest piece of fragility evidence in this program. It complements rather than duplicates the cross-model and repeated-sampling instability already documented: those show that which model and which sample matter more than conventional diagnostics suggest; this shows that the signal's specific content, holding everything else fixed, performs worse than most random relabelings of it.

One limitation of this extension should be weighed: the IG spread data reflects a snapshot of retail mini-contract pricing rather than institutional rates, a distinction that should be stated explicitly wherever cited. The permutation result itself now rests on the full planned 500 repetitions rather than a partial run, but it uses only independent reshuffling; block-preserving and cyclic-shift schemes test a different null hypothesis and have not been run here, so this result should not yet be read as robust to the choice of permutation scheme.