FinSentinel v2 @ FinMMEval 2026 Task 3: Online Threshold Calibration and Cross-Asset Regime Gating for Live Financial Decision Making

FinMMEval Task 3 at CLEF 2026

Awais Gill[1], Emaan Naveed Sheikh[1], Fatima Tu Zahra[1,*], Abdul Samad[1], Faisal Alvi[1] and Sandesh Kumar[1]

[1]*Habib University, Karachi, Pakistan*

Abstract

We present FinSentinel v2, our deterministic, rule-based trading agent submitted to CLEF 2026 FinMMEval Task 3 (Financial Decision Making). Building on a signal-fusion architecture with asset-specific decision policies, v2 extends our baseline entry, FinSentinel v1, with three targeted mechanisms: (1) **online threshold calibration**, which maintains a rolling 14-day accuracy buffer per asset and adapts BUY/SELL thresholds to counteract distributional shift between the training window and the live evaluation window; (2) a **cross-asset regime detector** that uses BTC short-term returns as a macro risk-on/off signal to modulate the TSLA BUY threshold; and (3) **asymmetric confidence weighting** that re-weights news and momentum signals proportionally to their recent per-signal accuracy. Over the 300-day live evaluation, FinSentinel v2 achieves a cumulative return (CR) of $+96.47\%$ and Sharpe ratio (SR) of 1.687 on BTC, with maximum drawdown (MDD) of 15.45% and action accuracy of 53.76%. We emphasise that this result is obtained during a period in which the buy-and-hold (BAH) baseline itself lost 35.05% of capital, so part of the raw margin over BAH reflects the bearish character of the evaluation window rather than signal quality alone; SR and MDD are the more direction-independent indicators of policy quality, and both improve substantially over our v1 baseline. On TSLA, the system achieves $+46.15\%$ CR (SR $= 1.453$, MDD $= 15.91\%$, Acc $= 42.86\%$) against a BAH of $+46.09\%$, matching passive return while roughly halving maximum drawdown relative to v1. All decision logic is deterministic; median inference latency is below 50 ms, well within the task's 3-minute endpoint timeout.

Keywords

financial decision making, trading agents, signal fusion, sentiment analysis, concept drift, cross-asset signals, live evaluation, CLEF 2026

1. Introduction

CLEF 2026 FinMMEval Task 3 requires an agent to issue a BUY, HOLD, or SELL decision on a fixed daily schedule over a live 300-day sequential evaluation period [1, 3]. Each decision fully replaces the prior position and is executed at market close; performance is assessed via risk-adjusted financial metrics—cumulative return (CR) and Sharpe ratio (SR) as primary measures, with maximum drawdown (MDD) and annualised daily volatility (DV) as secondary diagnostics. Classification accuracy against ground-truth labels is reported but is not the optimisation objective.

A note on versioning. Throughout this paper, "v1" refers to our team's initial baseline submission and "v2" to the extended system described here.

This evaluation protocol imposes three concrete design constraints. First, locally plausible decisions can compound into globally poor outcomes without explicit risk controls: a policy marginally above chance in classification terms can still produce severe drawdowns if errors concentrate at high-magnitude market moves. Second, the HOLD action retains the agent's current exposure and is also the response produced by endpoint failure or timeout, making it the operationally safe default under uncertainty.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

mg08993@st.habib.edu.pk (A. Gill); es09163@st.habib.edu.pk (E. N. Sheikh); ft09259@st.habib.edu.pk (F. T. Zahra); abdul.samad@sse.habib.edu.pk (A. Samad); faisal.alvi@sse.habib.edu.pk (F. Alvi); sandesh.kumar@sse.habib.edu.pk (S. Kumar)

Third, BTC and TSLA exhibit structurally different return regimes—BTC with higher volatility and stronger sensitivity to macro risk sentiment, TSLA with equity-specific drivers and trend persistence—precluding a single unified policy.

FinSentinel v1 deployed asset-specific rule-based policies over a lexicon-derived news sentiment score, price-derived technical signals, and a task-provided five-class momentum label. Post-hoc analysis of v1's live decision log, conducted by replaying decisions against the live price series and segmenting accuracy by volatility regime, identified two principal failure modes: (i) signal-threshold miscalibration arising from distributional shift between the window used to set static thresholds and the live evaluation window, most pronounced on BTC (v1 CR: $+1.51\%$, SR $= 0.081$); and (ii) TSLA policy insensitivity to macro risk-off regimes, which caused the agent to enter long positions at the onset of correlated equity drawdowns and contributed to an MDD of 32.91%.

FinSentinel v2 addresses both failure modes through three targeted mechanisms, all implemented deterministically without any LLM component at inference time:

- **Online threshold calibration** (Section 3.2): a per-asset rolling 14-day accuracy tracker that tightens BUY/SELL thresholds when recent non-HOLD decision accuracy falls below a defined floor and relaxes them when accuracy is high, implementing lightweight concept-drift adaptation [4] without model retraining.
- **Cross-asset regime detection** (Section 3.3): BTC short-term returns serve as a macro risk-on/off proxy for TSLA, exploiting documented co-movement between cryptocurrency markets and high-beta growth equities during risk-off episodes [5].
- **Asymmetric confidence weighting** (Section 3.4): news and momentum signals are re-weighted by their recent per-signal accuracy, down-weighting signals that have been systematically misleading in the most recent evaluation window.

As detailed in Table 1 and Table 3, FinSentinel v2 improves substantially over v1 on both CR/SR (BTC) and MDD/SR (TSLA). The central finding is that online calibration is the dominant performance driver on BTC, while cross-asset regime detection is the dominant risk driver on TSLA; neither effect requires model retraining or LLM-based reasoning at inference time. At the time of writing, the organisers had not yet released the cross-team leaderboard ranking for Task 3; we therefore do not report a placement figure here.

2. Related Work

2.1. Financial Sentiment Analysis

Loughran and McDonald [6] demonstrated that general-purpose sentiment lexicons systematically misclassify domain-specific financial language, motivating finance-specific word lists; earlier work by Antweiler and Frank [9] and Tetlock [10] similarly established that simple, transparent text-sentiment measures derived from financial discourse carry statistically detectable information about market behaviour, supporting the use of lexicon-based scoring as a lightweight but informative signal. Domain-adapted Transformer models such as FinBERT [7, 8] show strong gains on sentence-level sentiment benchmarks but function as point-in-time classifiers without sequential risk modelling or deterministic latency guarantees. Our lexicon-based sentiment module sacrifices recall depth for interpretability and guaranteed sub-50 ms inference, which the task's 3-minute endpoint timeout and live sequential protocol make operationally necessary.

2.2. Concept Drift and Online Threshold Adaptation

Gama et al. [4] survey concept-drift detection and adaptation strategies, establishing that static classifiers degrade predictably under distributional shift. Standard responses—detector-triggered retraining (ADWIN, DDM) and ensemble-based approaches—require either sufficient data for retraining or pre-registered model variants, neither of which is available under the task's live evaluation protocol. Our

online calibrator is a deliberate simplification: rather than retraining, it adapts decision thresholds in response to observed accuracy degradation over a 14-day rolling buffer. This is appropriate under the task's constraints, where the live evaluation provides no retraining opportunity and the signal space is low-dimensional enough that threshold adjustment constitutes a tractable first-order correction.

2.3. Cross-Asset Co-movement and Risk-Off Regimes

Bouri et al. [5] document significant co-movement between Bitcoin and broader equity markets during risk-off episodes. This motivates our use of BTC short-term returns as a macro risk proxy for the TSLA policy: within the task's multimodal payload, BTC signals are directly available, making them an operationally natural choice compared to external regime indicators such as the VIX. We make no causal claim; the rationale is correlational—contemporaneous BTC risk-off signals have historically co-occurred with correlated equity drawdowns sufficient to justify raising the TSLA BUY threshold.

2.4. LLM-Based Trading Agents

Qian et al. [3] find that agent architecture—specifically memory structure, decision coordination, and risk gating—is a stronger performance determinant than the choice of language-model backbone, and that memory-adaptive variants exhibit risk-averse behaviour in high-volatility regimes. Their work also frames LLM trading agents around a reasoning-to-action pipeline linking market perception to executable decisions. This motivates our design choice of encoding risk adaptation deterministically: the online calibrator and cross-asset regime gate implement analogous adaptive properties with lower latency and without dependence on model stochasticity or prompt sensitivity.

2.5. Task 3 in the FinMMEval Framework

FinMMEval [2] addresses limitations common to prior financial NLP benchmarks, including static backtesting protocols, single-language corpora, and synthetic news inputs. Task 3 instantiates a reasoning-to-action evaluation paradigm [3] with a hard 3-minute per-step response deadline and live sequential evaluation over 300 trading days on BTC and TSLA [1]. The live protocol means that, unlike backtesting, the evaluation window is fixed and unknown to participants at system design time—a constraint that directly motivates our focus on online adaptability rather than retrospective threshold optimisation.

3. System Description

FinSentinel v2 is deployed as a FastAPI service. At each daily step, the server receives a JSON payload specifying: target symbol, concatenated news text, a five-class momentum label (provided by the task evaluation infrastructure), current closing price, and date-stamped historical closing prices. Four modules process this payload in sequence; their outputs are consumed by an asset-specific decision policy. Figure 1 illustrates the architecture, including the feedback path by which observed decision outcomes update the online calibrator.

3.1. Signal Extraction

News Sentiment Score (ns). We maintain a manually curated lexicon of **128 phrase–weight pairs**: 45 general-domain positive phrases, 49 general-domain negative phrases, 14 TSLA-specific positive phrases, and 20 TSLA-specific negative phrases. The general-domain component covers cross-asset financial events (ETF approvals, earnings beats and misses, regulatory actions, bankruptcy filings); TSLA-specific extensions cover delivery-cycle terminology and margin-pressure language. Phrase matching is case-insensitive over the full concatenated news string. Weights range from 1 (generic

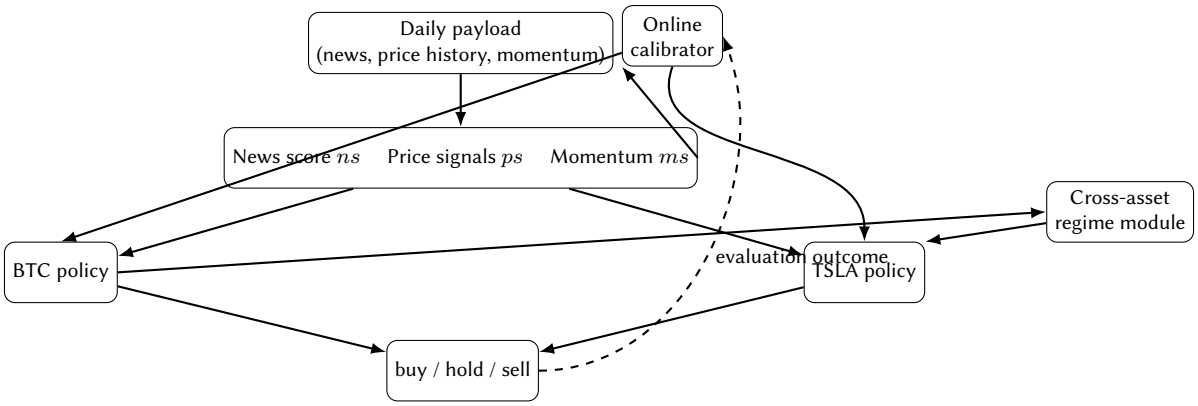

Figure 1: FinSentinel v2 system architecture. Solid arrows show the daily forward pass: signal extraction feeds both asset policies, the cross-asset regime module modulates the TSLA BUY threshold using live BTC return signals, and the online calibrator sets BUY/SELL thresholds for both policies. The dashed arrow shows the feedback path added in this revision: once the next day's price reveals whether a past non-HOLD decision was correct, that outcome is pushed back into the calibrator's rolling accuracy buffer, closing the online-adaptation loop described in Section 3.2.

directional language) to 5 (categorical high-impact events). The full lexicon is listed in Appendix A for reproducibility. The score is normalised to $[-1, +1]$:

$$ns = \frac{\sum_{p \in \text{POS}} w_p - \sum_{q \in \text{NEG}} w_q}{\sum_p w_p + \sum_q w_q}, \qquad ns = 0 \text{ if no phrases match.} \tag{1}$$

Price Signal Vector (ps). From the historical price series we compute: log returns over horizons r_3, r_7, r_{14}, r_{30}; 7-day realised volatility σ_7 (normalised against 30-day realised volatility σ_{30} for regime labelling); 20-day moving-average-relative position; 14-period RSI; and Bollinger Band percentile (BB%). Two composite indicators are derived: the dip signal $d = \min(1, (MA - p_t)/MA \times 10)$ and a peak signal, active only once BB% exceeds 0.85, defined as $\min(1, (\text{BB\%} - 0.85)/0.15 \times 5)$; the TSLA policy's peak-avoidance rule (Section 3.6) fires when this composite peak signal exceeds 0.4, which corresponds to BB% $\gtrsim 0.862$.

Momentum Score (ms). The task-provided five-class momentum label is mapped to a scalar: {strong bullish: $+1.0$, bullish: $+0.6$, neutral: 0.0, bearish: -0.6, strong bearish: -1.0}.

3.2. Novel Contribution 1: Online Threshold Calibration

Static threshold policies are sensitive to distributional shift between the window used to set thresholds and the live evaluation window. We introduce a per-asset `OnlineCalibrator` that maintains a rolling 14-day buffer of recent non-HOLD decision outcomes (correct/incorrect relative to next-day price direction). Let $\hat{a}_t$ denote the accuracy computed over this buffer. The calibrated BUY threshold is:

$$\tau_{\text{buy}} = \begin{cases} \min\left(0.80,\ \tau_0 + \dfrac{\alpha - \hat{a}_t}{\alpha} \cdot 0.15\right) & \hat{a}_t < \alpha \quad \text{(tighten)} \\ \max\left(0.35,\ \tau_0 - \dfrac{\hat{a}_t - 0.60}{0.40} \cdot 0.05\right) & \hat{a}_t > 0.60 \quad \text{(relax)} \\ \tau_0 & \text{otherwise} \end{cases} \tag{2}$$

where τ_0 is the asset-specific base threshold and $\alpha = 0.45$ is the accuracy floor below which tightening is triggered. An analogous formula governs the SELL threshold. The calibrator requires a minimum of 5 non-HOLD decisions before producing adjusted thresholds; during the warm-up period, τ_0 is used directly. When recent accuracy is poor, the system demands higher signal conviction before committing to a position; when accuracy is strong, it allows marginally more participation. This implements a form of lightweight concept-drift adaptation [4] without requiring model retraining or access to future data.

3.3. Novel Contribution 2: Cross-Asset Regime Detection

BTC is sensitive to macro risk sentiment: sharp BTC declines co-occur with risk-off episodes that also affect high-beta growth equities such as TSLA [5]. We maintain a `CrossAssetRegime` module updated each time a BTC decision step is processed. The normalised regime signal is:

$$\rho = \text{clip}\left(\frac{0.6 \cdot r_3^{\text{BTC}} + 0.4 \cdot r_7^{\text{BTC}}}{\max(3\sigma_7^{\text{BTC}}, 0.01)},\ -1,\ +1\right) \tag{3}$$

When $\rho < -0.4$ (risk-off), the TSLA BUY threshold is raised by 0.15 (capped at 0.75); when $\rho > 0.4$ (risk-on), it is lowered by 0.08 (floored at 0.20). Additionally, confirmed downtrend SELL signals are suppressed during risk-on periods to avoid premature exits into crypto-driven recoveries.

On the selection of $(0.6, 0.4)$ *and* $|\rho| = 0.4$. These constants were set by a manual, coarse parameter sweep over the v1 training period rather than an exhaustive or automated search, and we did not retain a full results table from that sweep. We flag this explicitly as a methodological limitation rather than reconstructing sensitivity results after the fact: the reported constants should be read as a reasonable, empirically motivated but not exhaustively validated configuration, and we discuss the consequence—a hard threshold at the regime boundary—in Section 5.4. We intend to release the calibration script and training-period logs alongside the source code so that a full sensitivity sweep can be reproduced independently.

3.4. Novel Contribution 3: Asymmetric Confidence Weighting

Each signal's contribution to the composite score is modulated by its recent per-signal accuracy, tracked within the calibrator buffer. Concretely, the effective weight of signal s at time t is scaled by $\hat{a}_t^{(s)}/\bar{a}_t$, where $\hat{a}_t^{(s)}$ is signal s's recent accuracy and $\bar{a}_t$ is the mean accuracy across signals, so that the weights sum to a constant. Signals with below-average recent accuracy are down-weighted; those with above-average accuracy receive proportionally more influence. This re-weighting is applied prior to the threshold comparison in both asset policies.

3.5. BTC Decision Policy

All non-HOLD decisions are gated behind minimum news-score thresholds supplied by the online calibrator. Conditions are evaluated in the following priority order: (1) strong-signal BUY: $ns \geq \tau_{\text{buy}}$ and $r_3 > -0.04 \Rightarrow$ buy; (2) strong-signal SELL: $ns \leq \tau_{\text{sell}}$ and $r_3 < 0.04 \Rightarrow$ sell; (3) moderate-signal with trend confirmation: $ns \geq \tau_{\text{buy}} - 0.05$ and $r_7 > 0.015 \Rightarrow$ buy, symmetric on the sell side; (4) low-volatility momentum, active only when $\sigma_7 < 0.018$, where a weighted composite of r_3, r_7, r_{14} must directionally agree with ms; (5) default: hold.

3.6. TSLA Decision Policy

The TSLA policy is long-biased and regime-conditional. Cross-asset regime modulation (Section 3.3) is applied before threshold evaluation. Conditions are evaluated in the following priority order:

1. **Confirmed downtrend SELL**: $r_{30} < -0.15$ and $r_7 < -0.03$ and price below the 20-day MA and RSI < 35 and not risk-on $\Rightarrow$ sell.
2. **Dip BUY**: $d > 0.25$ and RSI < 45 and $r_7 > -0.04$ and $r_{30} > -0.08$ and not risk-off $\Rightarrow$ buy, subject to the memory-gate veto below.
3. **Peak avoidance**: peak signal > 0.4 (BB% $\gtrsim 0.862$) and RSI $> 78 \Rightarrow$ hold.
4. **News-driven BUY**: $ns \geq \tau_{\text{buy}}$ (cross-asset adjusted) $\Rightarrow$ buy.
5. **Long-biased default** (suppressed in risk-off): buy if $ns > 0.30$, or $r_7 > 0.01$, or ($r_{30} > 0$ and RSI < 70); hold otherwise.

Table 1
Live 300-day evaluation results on the CLEF 2026 FinMMEval Task 3 benchmark. BAH = buy-and-hold baseline. MDD = maximum drawdown. DV = annualised daily volatility (%). Acc = action accuracy relative to ground-truth labels.

Asset	CR (%)	SR	MDD (%)	DV (%)	BAH CR (%)	Acc (%)
BTC	$+96.47$	1.687	15.45	2.28	-35.05	53.76
TSLA	$+46.15$	1.453	15.91	1.40	$+46.09$	42.86

Table 2
Action distribution over the 300-day live evaluation for FinSentinel v2. BTC BUY activity increased relative to v1 (158 vs. 111 days), consistent with the online calibrator relaxing thresholds during high-accuracy periods. TSLA issued zero SELL decisions; Section 4.3 discusses why the conjunctive SELL gate of Section 3.6 plausibly never fired during a broadly upward-trending TSLA window.

Asset	BUY (days)	HOLD (days)	SELL (days)
BTC	158	106	36
TSLA	240	60	0

Table 3
Performance comparison between FinSentinel v1 and v2 on the 300-day live evaluation. For MDD, a negative Δ indicates improvement (reduced drawdown).

Asset	System	CR (%)	SR	MDD (%)	Acc (%)
BTC	v1	$+1.51$	0.081	22.55	50.3
	v2	$+96.47$	1.687	15.45	53.8
	Δ	$+94.96$	$+1.606$	-7.10	$+3.5$
TSLA	v1	$+48.90$	1.145	32.91	38.8
	v2	$+46.15$	1.453	15.91	42.9
	Δ	-2.75	$+0.308$	-17.00	$+4.1$

Memory gate. A rolling 7-entry buffer records (action, 1-day return) pairs. A sell is vetoed if the prior buy yielded a return $r > +0.3\%$; a buy is vetoed if the prior sell yielded $r < -0.3\%$. This prevents the policy from reversing positions immediately after profitable entries.

We note explicitly that the SELL condition above is a five-term conjunction, all of which must hold simultaneously; this is a deliberately conservative gate given the policy's long bias, but it is also a plausible explanation for the zero SELL decisions observed on TSLA over the evaluation period (Table 2; discussed further in Section 4.3).

3.7. Endpoint and Operational Design

Signal computation is implemented in NumPy. The FastAPI handler is fully asynchronous; all error conditions return `{"recommended_action": "HOLD"}`, consistent with the task's timeout-as-hold semantics. A `/health` endpoint exposes real-time calibrator state (rolling accuracy estimate, current thresholds, cross-asset risk signal ρ) for operational monitoring. Median inference latency is below 50 ms on the serving hardware used during the live evaluation.

4. Experimental Results

4.1. Live Evaluation Results

Table 1 reports performance on the live 300-day evaluation. Table 2 shows the action distribution per asset. Table 3 compares FinSentinel v1 and v2 across all reported metrics; Table 4 summarises, per

component, which mechanism is attributable to which improvement.

4.2. BTC: Regime-Adaptive Outperformance

The BTC evaluation period was strongly bearish for passive investors; the BAH baseline lost 35.05% of capital. FinSentinel v2 returned $+96.47\%$ over the same period (Table 1), a 131.5 pp margin over BAH. We note that a declining market structurally favours any active strategy with a functioning SELL or HOLD capacity, so this margin partly reflects period-specific market directionality rather than signal quality alone. The more informative indicators of policy quality are SR $= 1.687$ and MDD $= 15.45\%$ (Table 1), which reflect the timing and risk-management properties of the calibrated thresholds rather than market direction, and both improve sharply over v1 (Table 3).

As Table 3 shows, the improvement over v1 is primarily attributable to the online calibrator enabling more active participation during periods of high signal accuracy. In v1, the static $\tau_{\text{buy}} = 0.50$ threshold suppressed many profitable entries during brief recoveries; the calibrator relaxed this threshold during high-accuracy windows, enabling 47 additional BUY days (158 vs. 111 in Table 2), while tightening it during signal breakdown to limit downside exposure. The resulting MDD reduction (22.55% $\rightarrow$ 15.45%) confirms that more active participation under calibration does not come at the cost of increased tail risk.

4.3. TSLA: Risk-Adjusted Parity with Substantially Reduced Drawdown

The TSLA evaluation period trended broadly upward ($+46.09\%$ BAH). In this regime, the relevant objective is not to outperform passive holding but to match it while reducing drawdown and improving risk-adjusted return. FinSentinel v2 achieved $+46.15\%$ CR and SR $= 1.453$, nearly matching BAH's absolute return with a substantially superior risk profile.

The most significant improvement is the MDD reduction from 32.91% (v1) to 15.91% (v2), a 17.0 pp decrease (Table 3). This is primarily attributable to the cross-asset regime detector: by suppressing BUY signals during BTC risk-off episodes ($\rho < -0.4$), the policy avoided initiating long positions at the onset of several market-wide drawdowns that constituted the majority of v1's maximum loss. The SR improvement (1.145 $\rightarrow$ 1.453) reflects more consistent daily returns: the cross-asset gate reduces the variance of BUY entry quality by filtering entries that coincide with macro risk-off conditions.

On the zero SELL decisions reported in Table 2: the SELL condition in Section 3.6 requires five conjunctive sub-conditions (a confirmed 30-day and 7-day downtrend, price below the 20-day moving average, RSI < 35, and the absence of a BTC risk-on signal) to hold simultaneously. Given that the evaluation window was broadly upward-trending for TSLA, it is plausible that this conjunction was never satisfied in practice. We view this as a genuine limitation of the current gate rather than a desirable property: a policy that cannot issue a SELL under any observed live-evaluation condition provides no empirical evidence, one way or the other, about whether its SELL logic is well-calibrated, and a less restrictive or graduated exit condition (Section 5.4) would let the policy demonstrate risk-reducing behaviour even in the trending regime actually observed.

Action accuracy of 42.86% is below the nominal three-class uniform baseline (33.33% would be chance, but the ground-truth label distribution is not uniform) while SR $= 1.453$ indicates strong risk-adjusted performance. This is consistent with a long-biased strategy that systematically diverges from ground-truth sell labels during drawdown periods when the policy is in hold or buy mode. As discussed in Section 5.3, accuracy is a limited proxy for policy quality in this evaluation setting.

4.4. Component Attribution

Table 4 consolidates the attribution discussed in Sections 4.2–4.3. Online calibration is the dominant driver of BTC improvement, enabling more active participation when signals are accurate while tightening exposure during signal breakdown. The cross-asset regime gate is the dominant driver of the TSLA MDD reduction, at a modest CR cost (-2.75 pp) reflecting some profitable BUY opportunities forgone during risk-off periods that did not ultimately result in drawdowns—the expected cost of a conservative regime filter. Asymmetric confidence weighting contributes smaller, secondary accuracy

Table 4
Component-level attribution summary, derived from decision-log analysis of the deployed v2 system (as in Table 3 and the discussion above): the primary observed effect associated with each mechanism, at the point in the decision pipeline where it is architecturally active.

Component	Asset	Primary observed effect
Online threshold calibration	BTC	$+47$ BUY days (111→158); CR $+94.96$ pp; SR $0.081 \rightarrow 1.687$; MDD $22.55\% \rightarrow 15.45\%$
Cross-asset regime gate	TSLA	MDD $32.91\% \rightarrow 15.91\%$ (-17.0 pp); CR -2.75 pp trade-off; SR $1.145 \rightarrow 1.453$
Asymmetric confidence weighting	Both	Acc $+3.5$ pp (BTC), $+4.1$ pp (TSLA); secondary effect on CR/SR relative to the two mechanisms above

gains and is most useful in ambiguous market conditions where multiple signals provide conflicting evidence.

5. Discussion

5.1. Online Calibration as Concept-Drift Adaptation

Section 3.2 and Table 4 together illustrate the practical cost of threshold miscalibration under distributional shift. The training window used to set v1's static thresholds encoded a specific volatility and news-signal correlation structure that did not persist throughout the 300-day evaluation. The online calibrator detected this shift through accuracy degradation and adjusted thresholds accordingly, without requiring access to future data or model retraining. This constitutes a practical instantiation of the lightweight adaptation strategies surveyed by Gama et al. [4], and suggests that in live sequential evaluation settings with fixed signal spaces, threshold adaptation may provide returns comparable to, or exceeding, those from model-capacity increases.

5.2. Cross-Asset Signals and Macro Risk Regimes

The TSLA MDD reduction (Table 3) demonstrates that cross-asset correlation is a tractable risk signal even in a simple deterministic system. BTC's high sensitivity to global risk sentiment makes it a natural proxy for growth-equity drawdown risk [5]. The modest CR trade-off relative to the MDD improvement and SR gain (Table 4) supports the hypothesis that regime-aware entry filtering provides a favourable risk-return modification for long-biased equity strategies in the evaluated period. The hard threshold ($|\rho| = 0.4$) is a known limitation, discussed further below; a continuous modulation of the BUY penalty as a smooth function of ρ would reduce the discontinuity at regime boundaries and may yield more consistent behaviour.

5.3. Action Accuracy as a Limited Evaluation Proxy

TSLA action accuracy (42.86%) is below the ground-truth label distribution baseline while SR $= 1.453$ indicates strong risk-adjusted performance. This reflects a well-documented property of asymmetric loss in financial decision making: a policy that concentrates errors on low-magnitude price moves while correctly handling high-magnitude moves can dominate, in return terms, a high-accuracy policy with uniformly distributed errors. This finding is consistent with Qian et al. [3], who observe similar decoupling between classification accuracy and financial performance metrics. Accuracy is reported for benchmark completeness but should not be interpreted as a proxy for policy quality under the task's risk-adjusted evaluation criteria.

5.4. Limitations and Future Work

We highlight five limitations relevant to interpreting these results and to future iterations of the system.

Position sizing. All BUY and SELL actions imply full capital allocation; fractional position sizing proportional to signal confidence would reduce MDD without a proportional reduction in expected CR.

Hard cross-asset threshold. The cross-asset regime threshold ($|\rho| = 0.4$) is discrete: TSLA's BUY threshold jumps rather than moving smoothly as ρ crosses the boundary. A continuous penalty function—for example, scaling the TSLA threshold adjustment proportionally to ρ itself rather than switching at a fixed cutoff—would give smoother risk adjustment at regime boundaries and is a natural next refinement to the design in Section 3.3.

Calibrator warm-up. The calibrator requires a minimum of 5 non-HOLD decisions before producing reliable accuracy estimates; a warm-start prior estimated from the training-period distribution would improve behaviour during the initial evaluation window.

Bearish-market context for BTC results. The BTC evaluation period—a bearish market with BAH $= -35.05\%$—is structurally favourable to active strategies, and the CR figure in particular should be interpreted with this context in mind; we consider SR and MDD the more portable indicators of policy quality across market regimes.

Overly restrictive TSLA SELL gate. As discussed in Section 4.3, the five-condition SELL conjunction issued zero SELL decisions over the full evaluation window. We regard this as an unverified rather than a validated design choice, and future revisions should either relax the conjunction, introduce an intermediate reduce-exposure action, or evaluate the gate explicitly against a held-out period containing a confirmed TSLA downtrend.

More broadly, evaluation across multiple market regimes—including bull markets and sideways consolidation periods, for both assets—would more robustly characterise system performance beyond the specific 300-day window reported here. We also note, for completeness, that the present ablation (Table 4) is attribution-by-analysis rather than independent leave-one-component-out ablation; obtaining the latter would require re-running the live evaluation with individual mechanisms disabled, which was not possible within the camera-ready period.

6. Conclusion

We presented FinSentinel v2, an asset-adaptive rule-based trading agent submitted to CLEF 2026 FinMMEval Task 3. The system extends the v1 architecture with online threshold calibration, cross-asset regime detection, and asymmetric confidence weighting—three mechanisms targeting distributional shift and macro risk-off insensitivity identified as failure modes through post-hoc v1 analysis.

On the 300-day live evaluation (Table 1), FinSentinel v2 substantially outperforms BAH on BTC and closely matches BAH on TSLA while cutting drawdown roughly in half relative to v1 on both assets (Table 3); the BTC margin should be read alongside the caveat that the period's bearish character mechanically favours active strategies. Online calibration is the dominant performance driver on BTC; cross-asset regime detection is the dominant risk driver on TSLA (Table 4). These results support the hypothesis that lightweight adaptive mechanisms targeting distribution shift and cross-asset correlation

provide reliable performance gains in live sequential financial evaluation without requiring increases in model complexity or the introduction of LLM-based reasoning at inference time. The refinements enumerated in Section 5.4—a graduated TSLA exit condition and a continuous cross-asset threshold in particular—mark clear directions for the next iteration of the system.

Acknowledgments

The authors thank the FinMMEval 2026 organisers for the task infrastructure and live evaluation environment.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools for limited language editing and camera-ready formatting assistance only. No generative AI tools were used to generate experimental results, conduct analyses, or produce the core scientific contributions of this work. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of this publication.

References

[1] Z. Xie, L. Qian, G. Georgiev, D. Dimitrov, R. Elbadry, et al., Overview of the FinMMEval 2026 Task 3: Financial Decision Making, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[2] Z. Xie, R. Elbadry, F. Zhang, G. Georgiev, X. Peng, L. Qian, J. Huang, D. Dimitrov, P. Nakov, et al., The CLEF-2026 FinMMEval lab: Multilingual and multimodal evaluation of financial AI systems, in: Advances in Information Retrieval, Lecture Notes in Computer Science, Springer, 2026. doi:10.1007/978-3-032-21321-1_37.

[3] L. Qian, X. Peng, Y. Wang, V. J. Zhang, H. He, H. Smith, Y. Han, Y. He, H. Li, Y. Cao, Y. Yu, A. Lopez-Lira, P. Lu, J.-Y. Nie, G. Xiong, J. Huang, S. Ananiadou, When agents trade: Live multi-market trading benchmark for LLM agents, arXiv preprint arXiv:2510.11695 (2025). doi:10.48550/arXiv.2510.11695.

[4] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation, ACM Computing Surveys 46 (2014) 44:1–44:37. doi:10.1145/2523813.

[5] E. Bouri, S. J. H. Shahzad, D. Roubaud, L. Kristoufek, B. Lucey, Bitcoin, gold, and commodities as safe havens for stocks: New insight through wavelet analysis, The Quarterly Review of Economics and Finance 77 (2020) 156–164. doi:10.1016/j.qref.2020.03.004.

[6] T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, The Journal of Finance 66 (2011) 35–65. doi:10.1111/j.1540-6261.2010.01625.x.

[7] D. Araci, FinBERT: Financial Sentiment Analysis with Pre-trained Language Models, Master's thesis, University of Amsterdam, 2019. arXiv preprint arXiv:1908.10063.

[8] Y. Yang, M. C. S. Uy, A. Huang, FinBERT: A pretrained language model for financial communications, arXiv preprint arXiv:2006.08097 (2020).

[9] W. Antweiler, M. Z. Frank, Is all that talk just noise? The information content of Internet stock message boards, The Journal of Finance 59 (2004) 1259–1294. doi:10.1111/j.1540-6261.2004.00662.x.

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

A. Sentiment Lexicon

Tables 5–8 list the full 128-entry phrase–weight lexicon used to compute the news sentiment score ns (Equation 1). **Note on the lexicon size:** an earlier draft of this paper stated "80 phrase–weight

pairs”; on reconciling the manuscript against the deployed system’s source, the lexicon in fact contains 128 entries (45 general-positive, 49 general-negative, 14 TSLA-positive, 20 TSLA-negative). We have corrected this figure throughout the camera-ready text (Section 3.1 and the abstract) to match the deployed system exactly, since the source code is the ground truth for what was actually evaluated live.

Table 5
General-domain positive phrases and weights.

Phrase	w	Phrase	w
beat expectations	4	record revenue	4
record profit	4	raises guidance	4
raised guidance	4	blowout earnings	5
etf approval	5	etf inflow	5
etf inflows	5	etf flows	4
etf approved	5	spot etf	5
institutional	3	inflow	2
approval	2	approved	2
partnership	2	acquisition	3
buyback	3	merger	3
outperform	3	buy rating	3
upgrade	3	adoption	3
breakthrough	3	surge	2
soar	2	rally	2
jump	2	rebound	2
recovery	2	gain	1
rise	1	growth	1
bullish	2	positive	1
increase	1	strong	1
halving	4	all time high	5
ath	4	strategic reserve	5
legal tender	4	sovereign	3
accumulation	2		

Table 6
General-domain negative phrases and weights.

Phrase	w	Phrase	w
bankruptcy	5	fraud	5
sec investigation	5	criminal	5
class action	4	delisted	5
default	5	short report	4
money laundering	5	seized	4
hack	5	breach	4
exploit	4	stolen	4
miss expectations	4	lowered guidance	4
cuts guidance	4	miss	2
disappointing	2	ban	4
crackdown	4	sanction	4
restrict	2	regulation	1
tariff	2	probe	3
investigation	2	short seller	3
downgrade	3	sell rating	3
underperform	3	crash	4
selloff	3	sell-off	3
outflow	3	dump	3
layoff	2	bearish	2
negative	1	weak	1
decline	1	drop	1
fall	1	loss	1
fine	2	penalty	3
warning	2	concern	1
fear	2		

Table 7
TSLA-specific positive phrases and weights.

Phrase	w	Phrase	w
record deliveries	5	delivery beat	5
deliveries beat	5	deliveries exceeded	4
production record	4	robotaxi	4
full self-driving	3	fsd approved	5
fsd launch	4	autonomous	3
optimus	3	megapack	2
gigafactory	2	cybertruck	2

Table 8
TSLA-specific negative phrases and weights.

Phrase	w	Phrase	w
price cut	4	price reduction	4
price war	3	cut prices	4
delivery miss	5	missed delivery	4
deliveries miss	5	delivery shortfall	4
demand concern	3	demand weak	3
weak demand	3	margin pressure	3
margin decline	3	margin compression	4
market share loss	3	byd	2
recall	3	musk sell	4
musk selling	4	discount	2