Working notes of TrustMe team in CLEF 2026 FinMMEval Track 3

Shamik Kundu[1,*], Akiko Kaneshiro[1] and Muneaki Imai[1]

[1]*Trust Co.,Ltd*

Abstract

This working note describes our submission to the CLEF 2026 FinMMEval Task 3 financial decision-making benchmark, in which an agent must produce a daily BUY / HOLD / SELL action for a stock and a cryptocurrency. Our system is a simple momentum-following agent built on a large language model: it follows the prevailing market trend by default and only steps aside when the evidence against the trend is clear. To arrive at this design we ran an extensive experimental program covering alternative trading strategies, different language-model backends, and a range of supporting modules; many of these alternatives looked attractive in isolation but failed under the constraints of the benchmark, in particular the fact that the supplied news describes events already reflected in the day's closing price. The note documents the deployed architecture, the experimental trajectory, and the lessons learned, including a sizeable catalog of negative results to give readers a complete picture of the path from initial ideas to the final agent. In the official live evaluation, the deployed agent returned $+3.44\%$ on TSLA and -8.15% on BTC; although BTC lost money in absolute terms, both assets out-performed a buy-and-hold benchmark by more than 18 percentage points in a falling market, reflecting the agent's long/flat, capital-preserving stance.

Keywords

LLM trading agents, prompt design, financial decision making, capital preservation, CLEF 2026

1. Task Description

1.1. Setup

CLEF 2026 FinMMEval [1] task 3 [2] is an online financial decision-making benchmark built on the When-Agents-Trade live multi-market evaluation framework [3]. Each day the participating endpoint receives a JSON observation containing: the asset's close price, a news payload (one or more articles describing same-day or recent events), a categorical momentum tag (`bullish` / `bearish` / `neutral`), and a 10-day price history. Within 180 seconds the endpoint must return a single action drawn from {BUY, HOLD, SELL}. BUY corresponds to taking a long position, SELL to a short position, and HOLD to flat; each action fully replaces the previous day's position.

The benchmark is scored primarily on Cumulative Return (CR), with Sharpe Ratio, Maximum Drawdown (MDD), and Volatility as secondary metrics. Two assets are evaluated: TSLA (equity, weekday-only) and BTC (cryptocurrency, 24/7). The official evaluation window spans from mid-May to late June / early July. In our internal evaluation we used four 31-day canonical sub-windows covering distinct regimes (Apr–May, strong bull, sideways, bearish) plus a number of early candidate windows used during exploratory ablation, stress windows, hard-day windows, and a best-30d window.

1.2. The lagging-news property

A defining structural property of this task is that the news provided in the daily observation is *lagging*: it describes events already reflected in the close price. We verified this empirically on the 229-day backtesting dataset.

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany

*Corresponding author.

shamik_kundu@trust-partner.co.jp (S. Kundu); akiko_kaneshiro@trust-partner.co.jp (A. Kaneshiro); muneaki_imai@trust-partner.co.jp (M. Imai)

0009-0009-0367-045X (S. Kundu)

Asset	News ↔ same-day move	News ↔ next-day move
TSLA	73.1%	51.2%
BTC	63.4%	50.3%

Table 1: News-price alignment on the backtesting dataset.

Same-day alignment is high (73%/63%); next-day alignment is at chance (51%/50%). This lagging-news structure is the single most consequential constraint on system design. It rules out any architecture that treats news as a predictive signal and explains why prior LLM-trading systems [4, 5, 6, 7, 3] – which assume real-time news streams – do not transfer directly. It also explains why the COLING-2025 crypto winners [8, 9] pivoted to sentiment classification plus rules rather than direction prediction.

1.3. Scoring and trial design

For each configuration we report Cumulative Return as the mean over up to three independent trials per window, together with the trial-to-trial spread (max − min). We report spread rather than standard deviation because with n=3 the spread is more interpretable. All backtest returns in this note are *gross*: our simulator applies no transaction costs and no slippage, and each action fully replaces the previous position. The official leaderboard, by contrast, applies transaction fees; we therefore report both the fee-adjusted and no-fee figures for the live run (Table 9), where the fee drag was small (≈0.5pp on TSLA, ≈0.7pp on BTC).

2. Model and System Architecture

2.1. Pipeline overview

The deployed system is organised around a runner which orchestrates four phases per day: (1) Analyze: ingest the observation and append a new row to the in-memory `PriceHistoryStore`; (2) Settle: patch yesterday's row with the realised return and generate the reflection summary; (3) Decide: run the LLM-based decision pipeline; (4) Persist: append the trade row and serialise the pipeline state.

The Decide phase composes four context-providing modules – News Analyzer (LLM), FinBERT sentiment (CPU), a 10-day sliding-window Memory, and dual-level Reflection – into a single trend-following prompt with a default-flip structure (Section 2.3). The system spans the full {BUY, HOLD, SELL} action space. BUY and HOLD are emitted by the LLM prompt; SELL is produced by a separate deterministic guard: before the LLM is called, a turbulence-index check on the recent price history (formally specified in Section 2.6) can override the decision and force a SELL on extreme-volatility days, bypassing the prompt entirely. This is the only path through which the system shorts. As we show in Section 2.6, we backtested this guard and found that shorting incurs asymmetric losses on this benchmark; we therefore ran the official submission with the guard held in reserve, so the live agent traded long/flat and emitted zero SELLs over the evaluation window (Section 5) — a deliberate, evidence-based configuration choice rather than an architectural limitation.

2.2. Active modules

News Analyzer. An LLM call that summarises the day's news payload into a structured object containing same-day sentiment, dominant narrative, and a horizon classification (one-day catalyst vs. multi-day theme). Used to inject a compact context into the trend-follow prompt.

FinBERT sentiment. FinBERT (ProsusAI, 110M params) [10] runs locally on CPU, scoring the first 20 sentences of each article. It returns a softmax over {positive, negative, neutral}. Critically, the prompt frames neutral FinBERT as "no bad news" rather than "no signal". Without this framing, FinBERT neutrality blocked virtually all trades.

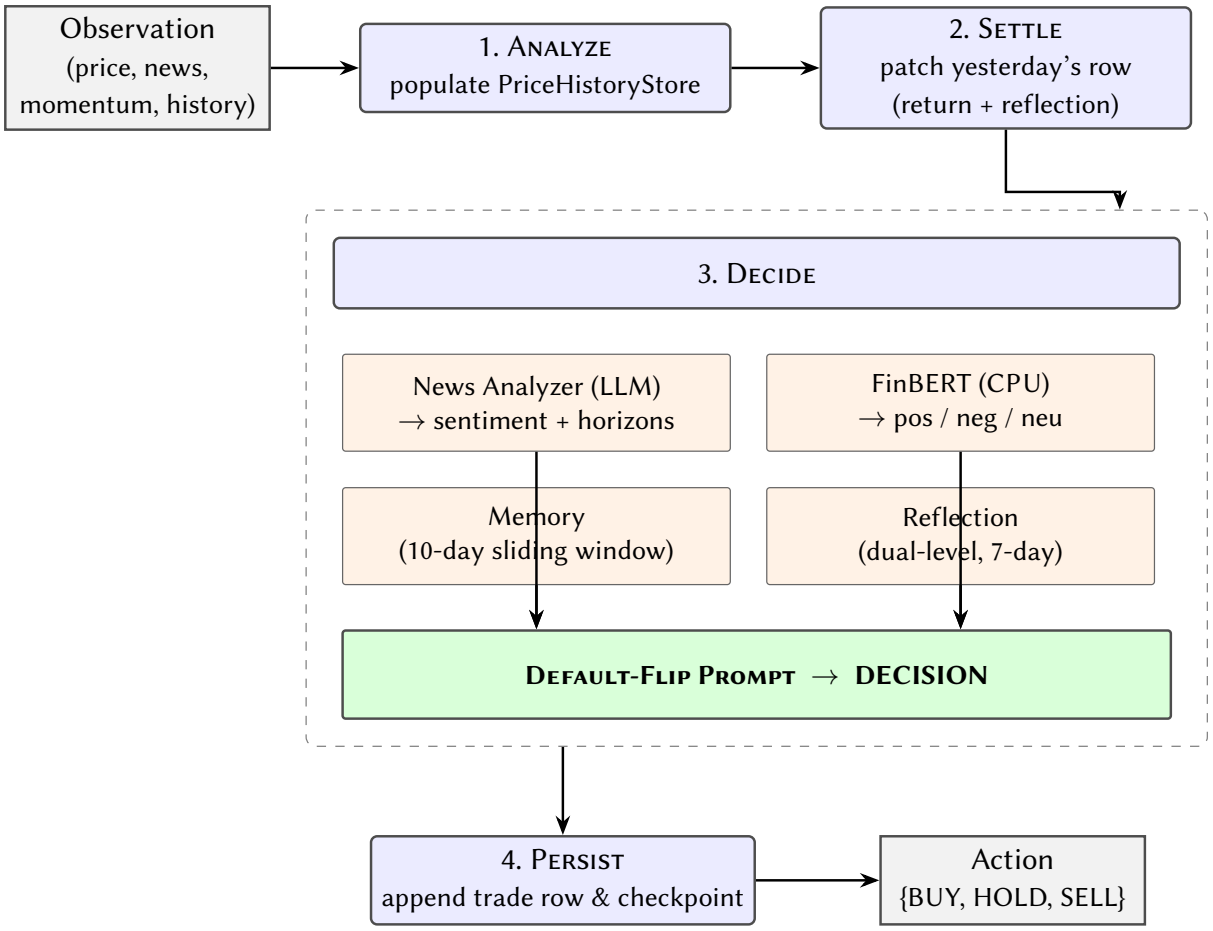

Figure 1: Deployed pipeline architecture. The daily observation flows through four phases (Analyze, Settle, Decide, Persist). The Decide phase fuses an LLM news analyser, a deterministic FinBERT sentiment signal, a 10-day memory window, and dual-level reflection into the default-flip prompt, which emits BUY or HOLD; SELL is produced by a separate deterministic turbulence-index guard that pre-empts the prompt on extreme down-days (Section 2.6, held in reserve for the live submission).

Memory. A 10-day sliding window of past decisions, prices, and outcomes is included in the prompt context. This is a deliberately small window: longer histories degraded performance (see ablation discussion).

Reflection. Dual-level: per-trade reflection writes a short post-mortem after each settled trade; a 7-day high-level summary is regenerated weekly. Both feed back into the decision prompt's context.

2.3. The default-flip prompt

The core design insight is the inversion of the LLM's default action. Conventional prompts in the LLM-trading literature use HOLD as the default, requiring "strong evidence" to trade but in a lagging-news setting such evidence never arrives. So we set the default:

- **BULLISH momentum** → default is BUY. Override to HOLD only if (i) today's price jumped >3%, (ii) news contains a clearly negative catalyst, or (iii) yesterday's BUY lost money.
- **BEARISH momentum** → default is HOLD. Exception: BUY if price dropped >3% with no new negative catalyst (oversold bounce).
- **NEUTRAL** → HOLD.

2.4. The prompt template

Figure 2 shows the static skeleton sent to the LLM each day, with the dynamically-filled slots highlighted in colour. The skeleton encodes the default-flip rules and the action constraint; every coloured slot is filled at runtime by the module shown beside it.

```
You are a stock trading analyst for {symbol} on {date}.
APPROACH: You follow the momentum. When momentum is BULLISH, you want
to be long. When momentum is BEARISH, you stay flat.
== TODAY'S MARKET STATE ==
Price: ${price}
Today's price change: {today_pct_label} from yesterday
{streak_text}
{regime_text}
FinBERT Sentiment: {finbert_label} (score: {finbert_net})
{decision_history_section}
== TODAY'S NEWS ==
{news_summary}
{memory_section}
{reflection_section}
== YOUR DECISION FRAMEWORK ==
Step 1 - MOMENTUM CHECK:
  - BULLISH → default BUY (need reason NOT to buy).
  - BEARISH → default HOLD (need strong reason to buy).
  - NEUTRAL → HOLD.
Step 2 - OVERRIDE CHECK (BULLISH only): HOLD only if today
  jumped >3%, news has negative catalyst, or yesterday's
  BUY lost money.
Step 3 - BEARISH EXCEPTION: BUY if price dropped >3% with
  no new negative catalyst and you did not buy yesterday.
== RULES ==
- You can only BUY or HOLD. Never SELL.
- Momentum is your primary signal. Trust it.
- Neutral FinBERT is fine for buying - it means "no bad news".
{asset_specific_rules}
{risk_instruction}
{mistake_section}
Output your confidence and decision on the LAST TWO LINES:
CONFIDENCE: 1|2|3|4|5
DECISION: BUY|HOLD
```

Market state (orchestrator + price store)

FinBERT sentiment

Decision history (recent BUY/HOLD outcomes)

News analysis (LLM-summarised, structured)

Memory (10-day sliding window of state)

Reflection (per-trade + 7-day high-level)

Asset rules / risk / mistakes (config-driven)

Figure 2: Default-flip prompt template. The static skeleton encodes the decision framework and action constraints; coloured slots are filled at runtime by the modules listed in the legend.

2.5. What is disabled and why

The deployed system disables several modules that were exhaustively tested:

- **Technical indicators (RSI/MACD/Z-score)**: the 10-day window is too short (RSI needs 14, MACD needs 26, Z-score needs 20). MACD repeatedly emitted "bearish" readings during clear uptrends because it was still anchored on stale prices.
- **Anti-inertia instruction**: gives +0.91pp mean over the deployed system but increases spread.
- **LLM-emitted SELL**: allowing the prompt to emit SELL directly produced asymmetric losses in shorts (the SELL-enabled overreaction baseline in Table 6; Section 6), so the LLM action space is restricted to BUY/HOLD and all shorting is routed through the deterministic turbulence guard below.

The turbulence-index SELL guard. The system's shorting path is a deterministic circuit-breaker on the price history. On each day t the guard computes a turbulence score $T_t = |r_t - \mu|/\sigma$, where r_t is the day's realised return and μ, σ are the mean and standard deviation of the prior returns over a 14-day trailing window (minimum 10 observations). It forces a SELL, bypassing the LLM, iff $T_t > \tau$ *and* $r_t < 0$,

where the threshold τ is the 90th percentile of the absolute z-scores of the prior-window returns; i.e. it shorts only on abnormally large down-moves. We backtested the guard by overlaying it on the deployed decision stream across the four windows. It fires on the extreme down-days (4 SELL days on TSLA, 6 on BTC across the four windows), but those days tend to mean-revert: enabling the guard lowers the four-window mean CR from $+3.52\%$ to -0.33% on TSLA and from $+4.96\%$ to $+4.25\%$ on BTC. This is the same asymmetric-short-loss pattern seen with LLM-emitted SELLs, and it is why we held the guard in reserve for the live submission (Section 5.5). The mechanism remains part of the architecture — it is the system's SELL path — but is disabled by default given this evidence.

2.6. Production deployment

The system is deployed on Google Cloud Run (2 vCPU / 4 GB, min-instances=1 to keep FinBERT warm). State is persisted to GCS. Typical end-to-end latency is 40–70 s per call, well within the 180 s SLA.

3. Training Setup

3.1. No model training

We do not train any model. All components are either pre-trained (FinBERT, the LLM backend) or hand-engineered (the prompts, the memory and reflection mechanisms, the default-flip decision rule). What we describe here as "training setup" is therefore really the experimental setup used to select among LLM backends, prompt variants, and module configurations.

3.2. LLM backends evaluated

We evaluated five LLM backends during development:

Backend	Role	Strength	Weakness
Gemini 2.0 Flash	Phase 1 ablation	cheap, fast	defaults to "mixed" sentiment
Gemini 2.5 Pro	Final deployed	strong reasoning	cost, latency 40–70 s
GLM-5	Strategy exploration	cheap, decent	poorer instruction following
DeepSeek R1	Alternative trial	3× faster, 4× cheaper	less stable on tails
DeepSeek v3.2	Spot checks	—	—

Table 2: LLM backends evaluated.

The deployed system uses Gemini 2.5 Pro. DeepSeek R1 with decision history was a competitive alternative but proved less stable on extreme-tail conditions.

3.3. Single-shot inference

All deployed and most reported results use single-shot inference with $n{=}1$ at temperature 0.1. Majority voting with three temperatures (0.1, 0.3, 0.5) was tested but it didn't meaningfully improve results; the deployed single-shot path retains only the first (lowest-temperature) sample.

3.4. Trial structure

For each (config, window) cell we run up to three independent trials. With $n{=}3$ we report spread (max − min) rather than standard deviation. Some exploration still used single-trial runs due to cost and deadline limitation.

3.5. The official submission

Our official submission: one Google Cloud Run deployment of the default-flip pipeline (single-shot $n{=}1$, Gemini 2.5 Pro), serving both TSLA and BTC from the same endpoint. It ran unattended for the full official window (2026-05-11 to 2026-06-26; 33 trading days for TSLA, 47 for BTC), and the numbers reported in Section 5 under "live/official evaluation" are the organiser-scored results for that one run. We did not submit multiple endpoints, nor did we ensemble or hand-pick among submitted runs.

3.6. Window-selection methodology

The four canonical windows (Apr–May, strong bull, sideways, bearish) were settled after exploratory ablation work on a set of early candidate windows. The goal was four 31-day windows that (a) span distinct regimes for both TSLA and BTC, (b) include the Apr–May window that overlaps with our live deployment, and (c) include at least one window where buy-and-hold loses money.

4. Features and Data Used

4.1. The daily observation

Each daily observation consists of:

- **Close price**: a single float for the asset.
- **News payload**: one or more articles.
- **Momentum tag**: categorical, {bullish, bearish, neutral}, *provided by the task organisers* as a field of the daily observation payload. It is not computed by our system; we consume it verbatim.
- **Ten-day price history**: closes for the trailing 10 trading days.

The momentum tag is the only forward-looking signal we trust; everything else is contemporaneous or lagging. The 10-day price history is too short for standard technical indicators, and we ultimately removed all indicator computations from the prompt context.

4.2. FinBERT signal

FinBERT [10] processes the first 20 sentences of each news article and emits softmax probabilities for {positive, negative, neutral}. We use this as a deterministic anchor: regardless of LLM stochasticity, the same article yields the same FinBERT vector. Crucially, the prompt presents FinBERT-neutral as "no bad news" rather than "no signal".

4.3. Momentum tag

The categorical momentum tag (bullish / bearish / neutral) is supplied by the task organisers as a field of each daily observation — it is task-provided, not computed by our system — and is rendered verbatim into the `{regime_text}` slot of the prompt (Figure 2); the default-flip decision framework then branches on it at every step. Because this tag is the only forward-looking signal we trust, the entire agent is, in effect, a momentum-conditioned policy over an organiser-provided signal — a point we return to when isolating the LLM's contribution (Section 5.3).

4.4. Ten-day price history

We use the 10-day window primarily for the LLM's contextual situational awareness: it is included in the prompt as a list of closes, not as a derived indicator. Internal experiments with extending to 30 days (to support RSI/MACD/Z-score) did not improve outcomes.

Window	Regime	TSLA B&H	BTC B&H
Apr–May (31 days)	Mixed	+7.98%	+17.28%
Strong bull	Bull	+33.2%	+5.3%
Sideways	Flat	−1.0%	−5.1%
Bearish	Bear	−6.1%	+2.3%

Table 3: Market-regime evaluation windows.

4.5. Dataset windows used for testing

We used four 31-days window with different characteristics for backtesting.

4.6. News characterisation

The news for both assets has the same structural property: every article contains both positive and negative keywords; "Overall sentiment" is hedged. The biggest BTC up-day had negative news ("market in distress"). LLM next-day direction accuracy from news alone is 27% on BTC – worse than a coin flip.

5. Evaluation Results

This section walks the reader from the headline results of the deployed system through the ablations that justify each design choice, the alternative strategies we tried and rejected. The narrative is thematic rather than strictly chronological.

5.1. Main results: the deployed default-flip pipeline

The deployed configuration is the reference against which all other experiments are compared. Tables 4 and 5 report three-trial means and spreads across all four canonical windows.

Window	T1	T2	T3	Mean	Spread
Apr–May	+7.4%	+8.1%	+9.9%	+8.45%	2.50pp
Strong bull	+13.0%	+11.2%	+11.2%	+11.81%	1.80pp
Sideways	−4.8%	−4.6%	−4.6%	−4.67%	0.20pp
Bearish	−2.5%	+1.2%	−1.6%	−0.96%	3.73pp
Average				**+3.66%**	**2.06pp**

Table 4: Deployed system on TSLA, 3 trials × 4 windows.

Window	T1	T2	T3	Mean	Spread
Apr–May	+5.6%	+8.6%	+8.6%	+7.59%	2.94pp
Strong bull	+4.3%	+2.1%	+4.3%	+3.53%	2.21pp
Sideways	+1.5%	−0.4%	+1.5%	+0.84%	1.89pp
Bearish	+1.2%	+0.8%	+3.1%	+1.68%	2.27pp
Average				**+3.41%**	**2.33pp**

Table 5: Deployed system on BTC, 3 trials × 4 windows.

The deployed configuration captures 35.6% of bull-market gains while protecting capital in bearish markets. Summed across the two assets, the four-window averages give +7.07% (+3.66% on TSLA and +3.41% on BTC).

5.2. Ablation matrix

Tables 6 and 7 place the deployed configuration alongside its closest variants. The salient finding is that components reduce trial-to-trial variance rather than improving mean returns.

Config	Description	Apr–May	Bull	Side	Bear	Avg CR	Spread	Avg Sh.	Avg MDD
Overreaction baseline	SELL enabled	+3.22%	+5.12%	−4.79%	−3.72%	−0.04%	—	—	—
Deployed	default-flip	+8.45%	+11.81%	−4.67%	−0.96%	+3.66%	2.06pp	1.71	−2.93%
+ anti-inertia	anti-inertia variant	+7.67%	+9.94%	−5.07%	−1.55%	+2.75%	1.98pp	0.81	−3.35%
+ technicals + risk	technical + risk variant	+5.90%	+11.17%	−3.55%	−1.18%	+3.09%	3.88pp	1.13	−2.78%
no history/memory/reflection	memory-ablated variant	+6.44%	+9.13%	−5.01%	+0.72%	+2.82%	2.59pp	2.07	−2.89%
no FinBERT	FinBERT-ablated variant	+6.58%	+11.89%	−4.60%	+0.34%	+3.55%	3.34pp	2.10	−3.02%

Table 6: All configurations – TSLA, mean CR and spread with mean Sharpe ratio (Avg Sh.) and mean maximum drawdown (Avg MDD) across the four windows (3 trials × 4 windows). Sharpe and MDD are averaged over the four windows. The overreaction baseline predates the structured-metrics logging and its secondary metrics are unavailable.

Config	Apr–May	Bull	Sideways	Bear	Avg Sh.	Avg MDD
Overreaction baseline	+5.99%	+0.00%	−6.89%	−1.24%	—	—
Deployed	+7.59%±2.94	+3.53%±2.21	+0.84%±1.89	+1.68%±2.27	2.09	−4.34%
+ anti-inertia	+4.42%±2.10	+2.71%±2.47	+0.29%±3.66	+2.17%±1.45	1.55	−4.77%
+ technicals + risk	+9.15%±1.76	+3.98%±0.43	+0.36%±3.31	+4.65%±5.53	2.64	−4.15%
no history/memory/reflection	+14.45%±12.90	+4.27%±3.10	+2.40%±2.05	+0.35%±1.92	—	—
no FinBERT	+8.57%±2.41	+5.69%±1.83	−0.93%±2.04	+3.10%±2.16	—	—

Table 7: All configurations – BTC, mean CR ± spread per window, with mean Sharpe (Avg Sh.) and mean maximum drawdown (Avg MDD) across the four windows. The memory- and FinBERT-ablated BTC variants were run only for their per-window CR and predate structured secondary-metric logging; their aggregate Sharpe/MDD are therefore not reported.

The overreaction baseline is dominated on every metric. The technical + risk variant raises BTC mean to +4.54% but doubles TSLA spread. Removing history/memory/reflection on BTC Apr–May delivers the highest single mean in the matrix (+14.45%) but at a trial-to-trial spread of ±12.90pp — larger than the mean itself. We rejected this configuration precisely because that variance makes it irreproducible: a single deployed run could as easily land near $+2\%$ as near $+14\%$, and we had no way to select the good draw in advance. This caution was vindicated by the live result, where BTC under-performed in absolute terms (Section 5.5); a high-variance configuration would have made that outcome even less predictable. The anti-inertia variant gives +0.91pp lower mean than the deployed system on TSLA at slightly lower spread; we kept the higher-mean configuration. The technical + risk variant trades a small TSLA loss for a BTC gain; we prioritised TSLA consistency. Across the matrix, the deployed configuration's distinguishing property is its tight spread, not the highest mean: components stabilise behaviour rather than lift it. The new secondary-metric columns (Avg Sharpe, Avg MDD) tell the same story — the deployed configuration is middling on mean Sharpe but never the worst on drawdown.

5.3. How much does the LLM add? A deterministic momentum baseline

Because the agent conditions heavily on the organiser-provided momentum tag (Section 4.3), we checked whether the LLM adds anything over a hand-coded rule. A fully deterministic momentum rule (BUY iff `bullish`, else HOLD, plus the two $\pm 3\%$ override conditions) achieves four-window mean CR of $+3.87\%$ on TSLA and $+4.97\%$ on BTC, versus $+3.66\%$ and $+3.41\%$ for the deployed LLM pipeline. In other words, the momentum tag carries essentially all of the directional signal, and the LLM does not improve headline return over a strong deterministic baseline. Its empirical value is instead (i) smoother behaviour on adverse windows — a naive "always buy the bull" rule collapses to -12.99% on TSLA sideways where the deployed agent limits the loss to -4.67% — and (ii) the ability to act on news catalysts, which a price-only rule cannot represent. We report this as an honest negative result and

regard the deterministic momentum rule as a baseline that any LLM agent on this benchmark should be measured against.

5.4. Things tested but not deployed

A long tail of components and decision mechanisms was exercised and discarded. We summarise them here.

Technical indicators. RSI/MACD/Z-score was tested but turned out to be not beneficial with 10-day history period.

Decision-mechanism variants. Three-way majority vote (temperatures 0.1, 0.3, 0.5) resulted in mostly HOLDs.

Dual-strategy and asset-specific prompts. A dual-strategy architecture with asset-specific prompts was implemented, motivated by the very different microstructure of stocks and crypto (Table 8). It was eventually replaced by a unified momentum-follow prompt due to comparable performance and lower maintenance burden.

Phenomenon	Stock	Crypto
Daily momentum corr.	∼50% neutral	36% anti-correlated
Mean reversion	moderate, multi-day	strong, same-day
Optimal RSI	standard 30/70	Cardwell 40–90 / 10–60
Oversold bounce	3+ down days	2+ down days >3%
Trading days	Mon–Fri	24/7
HOLD appropriateness	20–30%	<10%

Table 8: Stock vs. crypto market microstructure (dual-strategy motivation).

Volatility-adaptive BTC strategy. BTC mean-reversion works in high-volatility regimes (+21.85% on d187–216, 4.33% daily vol, 53% reversal rate) but fails in moderate-vol markets (−1.35% on d95–124, 2.55% daily vol, 41% reversal rate). A regime-switching design was prototyped but not used in final version.

5.5. Live / official evaluation

Table 9 reports the organizer-scored results for our official submission (Section 3.5) on the final frozen leaderboard.

Metric	TSLA	BTC
Cumulative return (with fees)	+3.44%	−8.15%
Cumulative return (no fees)	+3.94%	−7.54%
Vs. buy-&-hold	+18.26pp	+19.10pp
Sharpe ratio	2.76	−3.40
Win rate	75%	17%
Trading days	33	47
Leaderboard rank	5	15

Table 9: Final frozen official leaderboard results, evaluation window 2026-05-11 to 2026-06-26.

The headline of the live window is that it was a *falling* market for both assets: buy-&-hold lost heavily over the period. The deployed agent, being long-or-flat and sitting in HOLD on the majority of days,

protected capital and out-performed buy-&-hold by more than 18 percentage points on *both* assets. On TSLA it finished modestly positive (+3.44% with fees, Sharpe 2.76, 75% win rate). On BTC it finished negative in absolute terms (-8.15%), but still beat its buy-&-hold benchmark by +19.10pp. BTC is not a case of the agent destroying capital against a rising market, but of a hard down-market in which the agent lost far less than a passive holder.

BTC error breakdown. To locate the source of the BTC loss we decompose its 47 settled live days into their action types (the turbulence guard was held in reserve, so there were zero SELLs; the agent traded long/flat). The agent took 11 BUY days and 36 HOLD days. The loss is concentrated almost entirely in the BUY entries: only 2 of 11 BUYs were profitable (an 18% hit rate — consistent with the 17% leaderboard win rate), contributing -11.8pp in aggregate (-16.6pp across the nine losing BUYs against $+4.8$pp from the two winners). The 36 HOLD days, by contrast, were roughly a coin flip on the following day's direction (17 up vs. 18 down) and therefore neither helped nor hurt materially — but being flat on those days is exactly what kept the agent ahead of buy-&-hold. In other words, the BTC damage came from *mistimed BUY entries*, not from excessive inaction or forced SELLs. This is consistent with the lagging-news / chance-level next-day structure of BTC (Section 1.2): acting on the bullish momentum tag confers no reliable next-day edge on crypto, so BUYs behave like coin flips and, net of fees, bleed. It suggests that the BTC failure is structural to momentum-following on this 24/7 asset rather than an artifact of one window, and points to raising the bar for entering a BTC BUY (rather than enabling the turbulence SELL guard, which our backtest shows would have hurt — Section 2.6) as the natural next step.

6. Error Analysis

We document several negative results across our lab effort, organised below by root-cause category. The list is intentionally pruned – we keep only the most informative items in each category rather than enumerating every variant.

6.1. Lagging-news cause

1. **News-as-predictive-signal CoT.** $\sim$50% next-day accuracy. News describes same-day events; no forward edge.
2. **Sentiment-trajectory + TF-IDF novelty.** TSLA worst window degraded from -4.94% to -9.25%.
3. **Naive next-day prediction from FinBERT sign.** 51% TSLA, 50% BTC – chance.
4. **Overreaction framing for stocks.** 4–7 SELLs per window; lost on every window.
5. **Crypto-specific overreaction prompt.** 0 trades on strong-bull BTC over 31 days.

6.2. HOLD-loop and inaction biases

6. **HOLD-default prompt.** 0 trades produced.
7. **Self-reinforcing HOLD loop.** Prior HOLDs feed back as "no change" evidence.
8. **Unanimous-vote HOLD.** -5.51pp; suppresses nearly all trades.
9. **Three-way majority vote.** 62/62 TSLA calls returned HOLD; temperature diversity collapsed into conservatism.
10. **"Prefer HOLD when mixed" bias.** Compounded the inaction loop; removed.
11. **LLM confidence gating ($\geq$4).** The LLM never rates confidence ≥ 4; any threshold >3 produces all-HOLD (0/0/14 actions on TSLA d0–14 GLM-5 test).

6.3. Technical-indicator and rule-based failures

12. **Technical indicators (RSI/MACD/Z-score) on a 10-day window.** Noise from short windows; 30-day extension did not recover the balance.
13. **MA5/MA20 regime detection.** Required 20 prices, lagged ∼5 days.
14. **Regime hard-gate (no trades against MA regime).** −7.09% on TSLA d0–14; the MA regime was itself stale.
15. **Hurst exponent on 10-day windows.** Insufficient data points; −9.01pp vs baseline.
16. **Threshold sentiment override.** FinBERT scores not calibrated for rule firing.
17. **Pure rule-based price-structure trading.** +42% in-sample, 1/13 out-of-sample.
18. **Tech-Analyst sub-agent.** Dedicated technical-analyst LLM module was worse than no module.
19. **Threshold-hybrid (rule + LLM).** Worse than pure LLM; rules were already implicit in the prompt.

6.4. Strategy-family failures

20. **BTC mean reversion.** Edge in one window vanished on full dataset; 1-day and 2-day BTC mean reversion both 50/50.
21. **BTC volatility-adaptive strategy.** Worked in high-vol (+21.85%), failed in moderate-vol; not deployable.
22. **Always SELL on bear-market BTC.** +34.1% in hindsight, but timing-sensitive and impossible to implement online.
23. **Dual-strategy asset-specific prompts.** Comparable performance to unified prompt at higher maintenance cost.

7. Conclusion

The lab effort comprised 149 experiment runs across 24 experiment families, two assets and five LLM backends. The deployed default-flip pipeline is the result of systematic elimination of failure modes.

The most important findings of this lab effort are: (a) lagging news is a fundamental structural constraint that invalidates many architectures from the LLM-trading literature; (b) the default-flip prompt eliminates the self-reinforcing HOLD loop; (c) system components reduce trial-to-trial variance rather than improving mean returns; (d) LLM non-determinism is severe enough on crypto that single-shot inference with a deterministic anchor (FinBERT) outperforms majority voting; (e) a deterministic momentum rule over the organiser-provided tag matches or exceeds the full LLM pipeline's headline return (Section 5.3), so the LLM's contribution is downside smoothing and news-catalyst handling rather than higher mean return; and (f) in a falling live market the agent's long/flat stance out-performed buy-and-hold by more than 18 percentage points on both assets, even though BTC lost money in absolute terms — capital preservation, not directional alpha, is where the system delivered.

8. Data Usage Disclosure

The deployed system performs no dataset expansion and no real-time external retrieval at inference time. The only inputs consumed by the agent on each decision day are the fields provided in the task's daily observation payload (close price, news payload, momentum tag, 10-day price history). No web search, no live news API, no market-data feed, and no auxiliary dataset is queried during the official evaluation window. FinBERT is run locally on the supplied news text only. Model selection and prompt tuning were performed exclusively on the organiser-provided backtesting dataset; no other historical price or news data was used to train, prompt-engineer, or select among configurations.

9. Reproducibility

Software versions. Python 3.11; `transformers` 4.44.x with the `ProsusAI/finbert` checkpoint for FinBERT; Google `google-genai` SDK for Gemini 2.0 Flash and Gemini 2.5 Pro; the official Z.ai SDK for GLM-5; the DeepSeek OpenAI-compatible API for DeepSeek R1 and v3.2. The deployed single-shot pipeline issues its one decision call at temperature 0.1 (the three-temperature majority-vote experiment used 0.1, 0.3, 0.5).

Deployed configuration note. The deployed configuration wires a regime-aware risk module, but under the live 10-day price history it is effectively inert (it requires a 20-day moving average to change persona and therefore returns the neutral "moderate" persona, which injects no instruction into the prompt); the ablation experiments reported in Section 5 were run with the risk module disabled. The turbulence-index SELL guard (Section 2.6) is part of the architecture but was held in reserve (technical module off) for the official submission, following the backtest evidence that shorting hurts; the guard therefore did not fire during the live window. Readers reproducing the ablations should note that the anti-inertia instruction was removed from the trend-follow prompt in the deployed build; the anti-inertia ablation (Table 6, "+ anti-inertia") was produced from the prior build that still contained that instruction.

Random seeds. The deployed pipeline contains no explicit random sampling; FinBERT runs deterministically on CPU. The only source of stochasticity is the remote LLM endpoint, for which we do not control the seed. Reported means and spreads in Section 5 are computed across three independent end-to-end replays of the full window, which captures the API's effective non-determinism.

Hardware. Backtesting was run on a single workstation; the production endpoint is deployed on Google Cloud Run (2 vCPU, 4 GB RAM, `min-instances=1` to keep the FinBERT model warm in memory). FinBERT inference is CPU-only; all LLM inference is remote.

10. System Card

Architecture. See Section 2 and Figure 1: a four-phase runner (Analyze, Settle, Decide, Persist) in which the Decide phase fuses an LLM news analyser, a deterministic FinBERT sentiment signal, a 10-day memory window, and dual-level reflection into a single default-flip trend-following prompt that emits BUY or HOLD. The system spans the full {BUY, HOLD, SELL} action space: BUY/HOLD are LLM-emitted and SELL is produced by a deterministic turbulence-index guard (Section 2.6). Because backtests showed shorting incurs asymmetric losses on this benchmark, the guard was held in reserve for the official submission, so the live agent traded long/flat and emitted zero SELLs over the evaluation window.

Data utilization. See the Data Usage Disclosure above. No data beyond the daily observation payload is consumed at inference time.

Risk considerations. The agent is a research prototype for the FinMMEval Task 3 benchmark and is not intended for use as financial advice or for live trading of real capital. Known limitations: (i) absolute loss on BTC in live deployment (-8.15%, Table 9), traced to mistimed BUY entries in a chance-level next-day regime (Section 5.5) — note, however, that the agent still beat buy-and-hold by +19.10pp on BTC by remaining flat through most of the decline; (ii) the lagging-news property means the system cannot react ahead of price-moving information, and a deterministic momentum rule matches or exceeds the LLM's headline return (Section 5.3); (iii) remote LLM endpoints introduce non-deterministic and provider-dependent behaviour that cannot be fully audited. Outputs should not be interpreted as investment recommendations.

Acknowledgments

We thank the FinMMEval 2026 organisers for providing the Task 3 benchmark, the backtesting dataset, and the evaluation infrastructure. We acknowledge the use of FinBERT (ProsusAI) [10] as a sentiment-classification component, and the LLM API providers (Google, Z.ai, DeepSeek) whose services were used during development and deployment. All third-party components are used under their respective licenses; no proprietary or undisclosed data sources were used at evaluation time.

References

[1] Z. Xie, Y. Dai, R. Elbadry, V. Jani, X. Peng, L. Qian, G. Georgiev, D. Dimitrov, F. Zhang, J. Huang, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, S. Liu, P. Nakov, Overview of FinMMEval 2026: Multilingual and multimodal financial evaluation, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[2] Z. Xie, L. Qian, G. Georgiev, D. Dimitrov, R. Elbadry, F. Zhang, X. Peng, J. Huang, V. Jani, Y. Dai, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, S. Liu, P. Nakov, Overview of the FinMMEval 2026 task 3: Financial decision making, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[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, 2025. arXiv:2510.11695.

[4] Y. Yu, Z. Li, J. Zhang, Z. Gao, FINMEM: A performance-enhanced LLM trading agent with layered memory, in: AAAI Spring Symposium Series (AAAI-SS), 2024.

[5] W. Zhang, L. Jiang, et al., FinAgent: A multimodal foundation agent for financial trading, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2024.

[6] G. Fatouros, D. Metaxas, J. Soldatos, D. Kyriazis, MarketSenseAI 2.0, arXiv preprint arXiv:2502.00415 (2025).

[7] Y. Xiao, et al., TradingAgents: Multi-agents LLM financial trading framework, arXiv preprint arXiv:2412.20138 (2024).

[8] K. Lei, et al., FinNLP crypto trading challenge, in: Proceedings of the COLING FinNLP Workshop, 2025.

[9] Y.-J. Hsu, et al., Sam's fans at the crypto trading challenge, in: FinNLP Workshop at COLING, 2025.

[10] D. Araci, FinBERT: Financial sentiment analysis with pre-trained language models, arXiv preprint arXiv:1908.10063 (2019).