Infraglyph @ FinMMEval 2026 Task 3: EdgeQuant Agent – A Tiered Cognitive Memory Architecture*

Urvi Kava[1,*,†], Harsh Patel[1,*,†], Thomas Mandl[2,1] and Sandip Modha[1]

[1]Dhirubhai Ambani University, Gandhinagar, Gujarat, India
[2]Universität Hildesheim, Germany

Abstract

The proliferation of AI-driven trading agents for daily financial decision-making presents a fundamental challenge: avoiding the *HOLD bias* induced by risk-averse language generation while maintaining coherent temporal reasoning across market sessions. This paper describes the system developed by **EdgeQuant Agent** for Task 3 of the FinMMEval Lab at CLEF 2026, which requires daily BUY, HOLD, or SELL decisions for Bitcoin (BTC) and Tesla (TSLA) driven by real-time news, SEC filings, price history, and momentum signals. We propose a *four-tier cognitive memory architecture* integrating short-term news ingestion, mid-term SEC 10-Q quarterly filing context, long-term SEC 10-K fundamental regime anchoring, and a self-reflective decision layer. The temporal aspects are implemented with a ChromaDB vector store employing semantic retrieval using `BAAI/bge-small-en-v1.5` embeddings (384-dim). A supervised *warmup phase* (August 2025–April 2026, ≈270 trading days) uses an oracle for next-day price signals to build a corpus of justified reasoning traces. A subsequent *test phase* applies safeguards against lookahead shortcuts (date-limit filtering and mode-tag sanitisation) to prevent data leakage. Asset-specific expert personas and a *conviction-forcing mandate system* explicitly prohibit neutral HOLD decisions for active market sessions. On an evaluation window from January to April 2026 with GPT-OSS-120B as the reasoning backbone, the agent achieves a portfolio Sharpe Ratio of **3.165** versus an equal-weight benchmark of 2.387 (+32.6%), with a lower maximum drawdown of 8.22% (vs. 11.75%). The performance for TSLA reaches a Sharpe Ratio of $= 5.152$ and a Cumulative Return of $= 40.61\%$ over the evaluation window. On the official FinMMEval 2026 Task 3 Agent Market Arena leaderboard (BTC window: 2026-05-11 to 2026-07-04; TSLA window: 2026-05-11 to 2026-06-26), EdgeQuant Agent ranks **2nd** of 18 agents on TSLA (Return $+10.51\%$, Sharpe 3.34, Win Rate 67%, Final Balance $110,335.34) and **8th** of 18 agents on BTC (Return -0.52%, Sharpe 0.03, Win Rate 53%, Final Balance $99,316.51).

Keywords

financial decision making, large language models, retrieval-augmented generation, tiered memory architecture, autonomous trading agent, FinMMEval, CLEF 2026

1. Introduction

Autonomous financial decision-making using large language models (LLMs) has attracted growing research interest, with studies demonstrating that LLMs can encode and reason over financial news, sentiment signals, and macroeconomic narratives to inform trading decisions [1, 2]. However, naive LLM-based agents suffer from two well-documented failure modes: (1) a persistent *HOLD bias* arising from risk-averse language generation [3], and (2) *temporal amnesia* - the inability to consolidate and retrieve past market lessons from previous trading sessions [4].

We address both failure modes in **EdgeQuant Agent**, an autonomous trading system designed for the FinMMEval 2026 benchmark [5], specifically for Task 3 on financial decision making [6]. The evaluation setup follows the live multi-market trading framework introduced in *When Agents Trade* [7]. The system introduces a *four-tier cognitive memory architecture* that partitions market knowledge by temporal

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

* Code and configuration available at: https://github.com/Urvikawa31/EdgeQuant-Agent

*Corresponding author.

†These authors contributed equally.

urvikawa2004@gmail.com (U. Kava); harshpatel080503@gmail.com (H. Patel); mandl@uni-hildesheim.de (T. Mandl); sandip_modha@dau.ac.in (S. Modha)

0009-0009-5418-0768 (U. Kava); 0009-0007-5694-6486 (H. Patel); 0000-0002-8398-9699 (T. Mandl); 0000-0003-2427-2433 (S. Modha)

horizon—mirroring how professional fund managers process information across intraday, quarterly, and long-run fundamental timescales — backed by a ChromaDB vector store [8] for persistent semantic retrieval, enabling the agent to access contextually relevant past knowledge before each trading decision. A **conviction-forcing mandate system** explicitly prohibits HOLD decisions during active trading sessions, requiring the LLM to commit to directional positions based on available evidence.

Contributions.

1. A **four-tier cognitive memory architecture** (short / mid / long / reflection) backed by ChromaDB with semantic retrieval, temporal decay scoring, and anti-lookahead guards for clean train–test separation.
2. An **asset-specialised expert persona system** that injects domain-specific analytical priors (BTC: liquidity/ETF focus; TSLA: delivery variance/margin focus) without requiring model fine-tuning.
3. A **conviction-forcing mandate system** with asset-adaptive and calendar-adaptive rules that eliminates the HOLD bias for active market sessions.
4. Comprehensive **backtesting across four LLM backends** (GPT-OSS-120B, Mistral-large-3:675B, Gemma3-27B, Nemotron-Super), isolating the contribution of model capability from architectural design.

2. Related Work

Financial decision-making systems have evolved from traditional sentiment-analysis pipelines toward autonomous reasoning agents powered by large language models (LLMs). Early financial NLP research focused on extracting sentiment and market signals from news and social media using conventional machine learning methods [9]. With the rise of transformer-based architectures, finance-specialised language models such as BloombergGPT [10], FinGPT [11], and FinMA [12] demonstrated stronger financial reasoning and domain adaptation capabilities. Subsequent studies explored whether LLMs could directly support trading decisions. Lopez-Lira and Tang [1] showed that ChatGPT-generated sentiment exhibits measurable return predictability, while Koa et al. [2] investigated prompt-based stock selection using generative models. More recent systems such as FinMem [13], FinAgent [14], and TRADER [15] introduced memory-enhanced and tool-augmented trading agents capable of multi-step financial reasoning. In parallel, retrieval-augmented generation (RAG) [16] and cognitive agent frameworks such as Generative Agents [17] and Reflexion [18] demonstrated that layered memory and self-reflective reasoning can improve long-horizon decision making. EdgeQuant Agent builds upon these developments by integrating hierarchical memory retrieval, reflection-driven reasoning, and conviction-oriented decision making within a unified trading architecture.

Recent work has also focused on evaluating financial reasoning capabilities under realistic market conditions. InvestorBench [19] and FLARE-based evaluations [20] benchmark the performance of modern LLMs across financial reasoning tasks, while broader studies on foundation models in finance [4] highlight the growing importance of multimodal reasoning, risk-aware inference, and agentic financial systems. Within this direction, FinMMEval 2026 [5] introduced a multilingual and multimodal benchmark for financial AI evaluation, and Task 3 [6] extends this setting into daily trading decision making using evolving market information streams. EdgeQuant Agent is designed within this emerging research landscape and focuses specifically on temporally grounded memory retrieval, anti-lookahead reasoning safeguards, and risk-sensitive autonomous trading.

3. Task Description

FinMMEval 2026 [5] Task 3 [6] is formulated as a *daily, news-driven trading workflow*. Each submitted endpoint receives one HTTP POST request per day and must return BUY, HOLD, or SELL. The payload includes the trading date, current price, recent news headlines, a momentum label (bullish / bearish / neutral), a 10-day historical price series, and optionally SEC 10-K and 10-Q filing

summaries. The returned action maps to long flat, or short positions at daily close price; failed requests default to HOLD within a 3-minute timeout.

Assets. The task focuses on two assets: **Bitcoin** (BTC), a highly volatile digital cryptocurrency influenced by macro liquidity conditions and institutional ETF flows; and **Tesla** (TSLA), a high-beta equity shaped by vehicle delivery performance, autonomous driving developments, and competitive dynamics within the EV industry. Throughout this paper, the symbols BTC and TSLA will be used to refer to Bitcoin and Tesla, respectively.

Evaluation Metrics. The primary metric is Cumulative Return (CR). Secondary metrics include Sharpe Ratio (SR), Maximum Drawdown (MD), Daily Volatility (DV), and Annualised Volatility (AV).

Dataset. Historical training data was sourced from the `MBZUAI/finmmevallabclef2026` HuggingFace collection, specifically the `TheFinAI/CLEF_Task3_Trading` split. The dataset provides daily records from August 2025 to April 2026 for both BTC and TSLA, including pre-summarised multi-article news digests (synthesised by a separate LLM preprocessing pipeline), daily close prices, momentum indicators, and SEC filing summaries where available.

4. System Architecture

EdgeQuant Agent consists of three tightly integrated components: Data ingestion and pre-processing, the cognitive memory system, the LLM-based reasoning engine. Figure 1 illustrates the end-to-end pipeline through which market information is transformed into retrieval-grounded trading decisions.

4.1. Data Ingestion and Pre-processing

The agent operates across three sequential phases: *warmup*, *test*, and *evaluation*. During warmup, the model receives access to next-day price changes in order to generate grounded reasoning traces that are persistently stored in memory. Test mode removes all future information and enforces strict temporal filtering to prevent warmup-generated reflections from influencing live decisions. Evaluation computes cumulative return, Sharpe Ratio, maximum drawdown, and annualised volatility against both the agent portfolio and an equal-weight benchmark.

At each trading step, the market environment ingests date-keyed JSON records containing pre-summarised news digests, SEC filings, historical price windows, and momentum signals for BTC and TSLA. These signals are processed chronologically to construct a unified market-state representation. Momentum is modelled as a 3-day cumulative directional signal:

$$m_t = \text{sign}\left(\sum_{i=1}^{3}(p_{t-i+1} - p_{t-i})\right) \in \{\text{bullish},\ \text{neutral},\ \text{bearish}\} \tag{1}$$

All ingestion components construct a unified feature vector, serving as the standardised market-state representation for downstream reasoning and execution:

`date` | `symbol` | `closing_price` | `history_price` (10-day) | `news` | `10k` | `10q` | `momentum` | `future_price_diff`

The `future_price_diff` field is populated with the realised next-day price change only during the warmup phase, where it is passed to the LLM to construct oracle-informed reasoning traces. In test mode, this field is explicitly withheld (set to `None`) before prompt construction, so that no future price information ever reaches the reasoning engine during live decision-making; this feature-level safeguard operates in addition to the memory-retrieval-level anti-lookahead filtering described in Section 4.2.

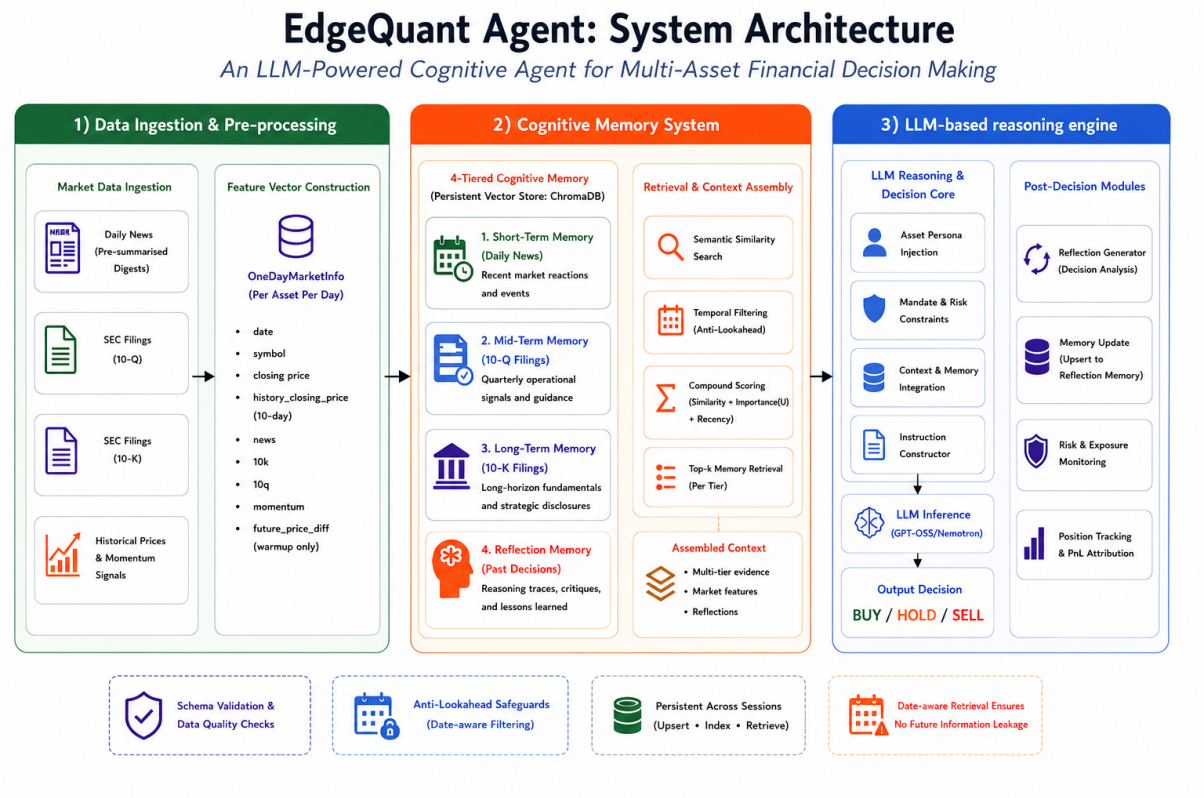

Figure 1: System architecture of **EdgeQuant Agent**. The framework integrates market data ingestion, hierarchical memory retrieval, and LLM-driven financial reasoning for autonomous multi-asset trading. Financial news, SEC filings, and historical market signals are transformed into structured feature vectors and used to generate BUY, HOLD, or SELL decisions through time-aware retrieval with anti-lookahead safeguards.
Source: Workflow conceptualised by the authors; figure visually generated using Google Gemini AI.

Schema validation is enforced during ingestion;malformed or incomplete records trigger a structured exception that is propagated to the competition API fallback handler, which defaults the response to HOLD.

The cognitive agent core orchestrates the complete decision pipeline through a unified `step(market_info)` interface. At each step, the agent retrieves semantically relevant memories from all four ChromaDB layers, assembles a retrieval-aware prompt, performs LLM inference, parses the generated decision, and updates the reflection memory with the resulting reasoning trace. Structured validation and fallback mechanisms ensure that malformed outputs or runtime exceptions default safely to HOLD decisions rather than system failure.

To support reproducibility and post-hoc analysis, EdgeQuant Agent implements a lightweight observability layer that records structured trade logs, retrieval metadata, inference latency, reasoning traces, and portfolio statistics throughout warmup and evaluation. These logs enable detailed analysis of memory utilisation, retrieval relevance, and model behaviour across trading sessions.

4.2. Tiered Cognitive Memory System

Inspired by cognitive memory models from generative-agent research [17], EdgeQuant Agent implements a **four-tier memory architecture** that partitions market knowledge across multiple temporal horizons. Short-term memory stores recent news events, mid-term memory captures quarterly operational signals from SEC 10-Q filings, long-term memory maintains strategic information extracted from annual 10-K filings, and reflection memory stores past trading decisions together with their reasoning traces. Figure 2 illustrates the complete memory architecture, while Table 1 summarises the

configuration of each layer.

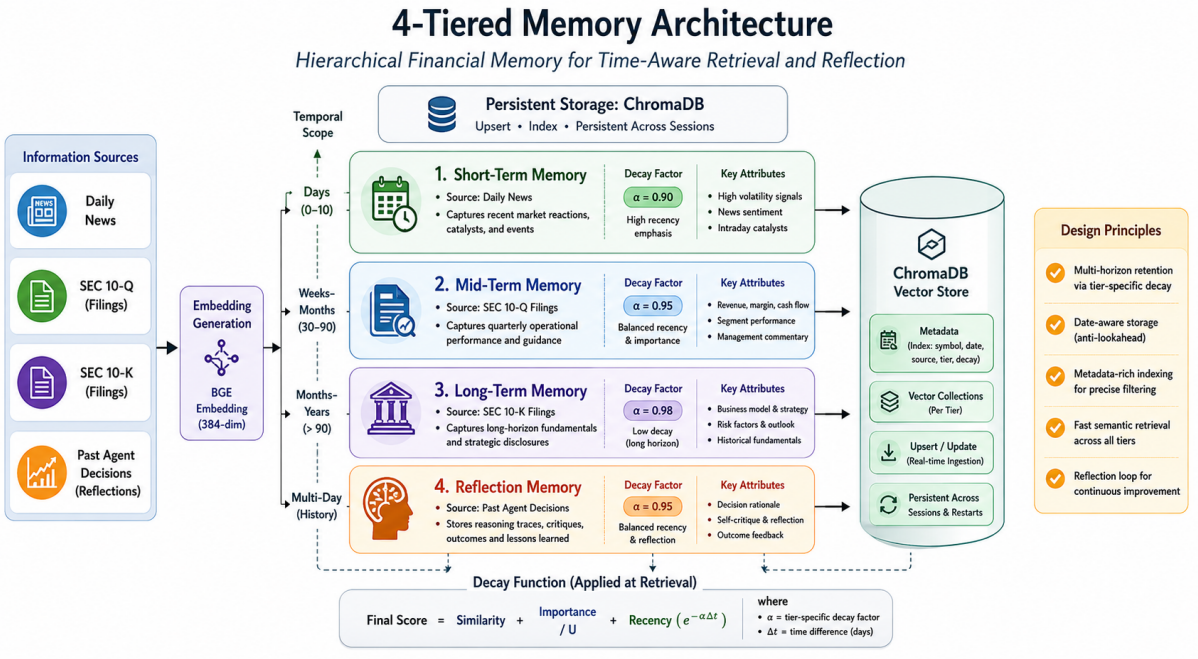

Figure 2: The 4-tier cognitive memory architecture stores market knowledge across multiple temporal horizons and enables retrieval-grounded financial reasoning using contextually relevant and self-reflective evidence. *Source: Architecture conceptualised by the authors; figure visually generated using Google Gemini AI.*

Table 1
Memory layer configuration.

Layer	Source	Decay α	Recency τ	Init. Importance
Short	Daily news digests	0.90	10	1.0
Mid	SEC 10-Q filings	0.95	20	2.0
Long	SEC 10-K filings	0.98	50	3.0
Reflection	Past agent decisions	0.95	20	2.0

All memories are stored in ChromaDB [8] as persistent vector embeddings generated using the `BAAI/bge-small-en-v1.5` model [21]. Each record stores the asset symbol, timestamp, memory tier, textual content, and execution mode (warmup/test), enabling efficient metadata-based temporal filtering. During inference, the agent queries all four memory layers using the current day's news digest as the retrieval query. Retrieved entries are ranked using a compound retrieval score combining semantic similarity, accumulated memory importance, and temporal recency:

$$s = s_{\text{sim}} + \frac{\min(s_{\text{imp}}, U)}{U} + s_{\text{rec}} \tag{2}$$

where s_{sim} denotes cosine similarity, s_{imp} represents accumulated memory importance, and $s_{\text{rec}} = \exp(-\Delta/\tau)$ models temporal decay. The top-k memories from each layer are injected into the final LLM context after ranking and pruning.

To prevent information leakage during evaluation, all retrieval operations apply strict anti-lookahead safeguards. Test-mode queries enforce `date_int < current_date_int` filtering, ensuring that no future information is accessible during inference. In addition, reflection memories generated during warmup are tagged with `mode=warmup` and excluded from test-time retrieval to prevent warmup-generated reasoning traces from influencing production decisions. After each trading step, the agent

generates a structured reflection summarising its decision rationale, market interpretation, and conviction level, which is embedded and stored back into the reflection layer. Over time, this evolving corpus of trade journals enables retrieval of semantically similar historical reasoning patterns, implementing a lightweight form of verbal reinforcement learning [18].

4.3. LLM-based Reasoning and Decision Framework

The EdgeQuant agent employs asset-specialized prompting strategies to align the LLM's reasoning process with the structural characteristics of different financial instruments. Rather than relying on generic financial prompts, the system injects expert personas and asset-specific analytical policies into the reasoning context. The BTC agent is configured as a global macro digital assets analyst focused on liquidity flows, institutional ETF activity, and market structure signals, while the TSLA agent follows a skeptical equity research framework emphasising delivery variance, competitive pressure, margin compression, and expectation surprises. These policies guide the model toward financially grounded reasoning without requiring model fine-tuning.

For TSLA, the reasoning framework evaluates catalysts according to freshness, magnitude, expectation variance, and temporal relevance, distinguishing between long-horizon structural developments and short-lived market noise. For BTC, the reasoning policy prioritises institutional liquidity flows, custody activity, ETF demand, and on-chain confidence signals while explicitly suppressing retail-driven hype indicators. Together, these constraints reduce spurious directional commitments and improve the quality of generated reasoning traces.

Before prompt injection, retrieved memories are ranked and pruned using layer-specific context budgets in order to prevent context-window overflow. Higher-density sources such as SEC 10-K filings receive larger token allocations, while news digests and reflection memories are truncated more aggressively. This hierarchical pruning strategy preserves the most temporally and semantically relevant evidence while maintaining compatibility with standard LLM context limits.

To mitigate the strong neutral bias commonly observed in financial LLM reasoning, EdgeQuant Agent applies a conviction-forcing mandate that encourages directional decisions under sufficiently strong evidence. BTC prompts prohibit HOLD decisions due to the continuous 24/7 nature of crypto markets, whereas TSLA permits HOLD decisions only during low-information weekend periods. Final LLM responses are parsed through a structured multi-stage extraction pipeline, and any unrecoverable parsing failure defaults conservatively to a HOLD decision.

Following inference, the post-decision module updates the reflection memory with the generated reasoning trace, records trading statistics, and performs lightweight risk and exposure monitoring. These post-execution updates enable persistent cross-session learning while supporting downstream evaluation and behavioural analysis across trading sessions.

5. Experiments and Results

5.1. Experimental Setup

Data. The `TheFinAI/CLEF_Task3_Trading` dataset was loaded from HuggingFace and converted to date-keyed JSON files. The dataset spans approximately 270 trading days (August 2025–April 2026) for both assets, including pre-aggregated news summaries, price data, momentum labels, and SEC filing summaries.

Warmup Configuration. LLM temperature: 0.2; memory retrieval top-k: 5 per layer; momentum window: 3 days; history window: 10 days; checkpoint saved after every step.

Sharpe Ratio Computation. For the backtest results in Tables 2–4, the Sharpe Ratio is computed on daily portfolio returns as $\mathrm{SR} = (\bar{r}/\sigma_r)\sqrt{252}$, where $\bar{r}$ and σ_r are the mean and standard deviation of daily returns and the risk-free rate is assumed to be $r_f = 0$. This $\sqrt{252}$ annualisation factor is

applied uniformly to both TSLA and BTC; we note this is an approximation for BTC, which trades on a continuous 365-day calendar, and our backtest does not apply an asset-specific trading-day adjustment. The official leaderboard results in Table 5 are computed independently by the FinMMEval 2026 Agent Market Arena platform using its own internal scoring methodology; the exact risk-free rate and annualisation convention used by the official platform are not documented in the competition materials available to us, and we report those values as provided by the leaderboard.

Models Evaluated. We evaluated four cloud-hosted LLM backends using the same memory checkpoint and prompt templates. Performance differences are therefore attributable to model reasoning capability:

1. `gpt-oss:120b-cloud` (GPT-OSS 120B) — Primary submission model
2. `mistral-large-3:675b-cloud` (Mistral Large 3, 675B)
3. `gemma3:27b-cloud` (Gemma3, 27B)
4. `nemotron-3-super:cloud` (NVIDIA Nemotron Super)

5.2. Portfolio-Level Results

Table 2 presents portfolio-level metrics across all evaluated LLM backends, compared against the equal-weight benchmark (50% BTC + 50% TSLA, rebalanced daily).

Table 2
Portfolio performance across LLM backends (January–April 2026) over the equal-weight benchmark.

System	Cum. Return	Sharpe Ratio	Max Drawdown	Ann. Volatility
Equal-Weight Benchmark	0.1886	2.387	0.1175	0.3816
GPT-OSS-120B	**0.1809**	**3.165**	**0.0822**	**0.2662**
Mistral-large-3:675B	0.0436	0.904	0.1160	0.2690
Gemma3-27B	0.0158	0.429	0.1012	0.2493
Nemotron-Super	−0.0544	−0.867	0.1348	0.2703

The GPT-OSS-120B configuration achieves a Sharpe Ratio of 3.165 ($+32.6\%$ over the benchmark SR of 2.387), a maximum drawdown of only 8.22% (vs. 11.75% benchmark), and annualised volatility of 26.6% (vs. 38.2%), demonstrating successful volatility dampening through selective directional positioning.

5.3. Per-Asset Results

Table 3 reports individual asset metrics for the GPT-OSS-120B configuration.

Table 3
Per-asset performance, GPT-OSS-120B, January–April 2026.

Asset	Cum. Return	Sharpe Ratio	Max Drawdown	Volatility
TSLA	**0.4061**	**5.152**	**0.0949**	0.3323
BTC	−0.0444	−0.374	0.1910	0.3871

TSLA emerges as the dominant alpha source, delivering exceptional risk-adjusted performance (SR $= 5.152$, CR $= 40.61\%$). BTC underperforms across all tested models, consistently generating negative cumulative returns.

5.4. Model Sensitivity

Table 4 decomposes per-asset performance across all four model backends, illustrating the extent of model dependency.

Table 4
Per-asset, per-model performance decomposition (January–April 2026).

Model	TSLA CR	TSLA SR	BTC CR	BTC SR
GPT-OSS-120B	**0.4061**	**5.152**	−0.0444	−0.374
Mistral-large-3:675B	0.2143	2.900	−0.1272	−1.661
Gemma3-27B	0.1322	1.910	−0.1007	−1.137
Nemotron-Super	0.0440	0.772	−0.1527	−1.894

5.5. Official Live Leaderboard Performance

Table 5 reports the official per-asset results on the FinMMEval 2026 Task 3 Agent Market Arena leaderboard.

Table 5
Official FinMMEval 2026 Task 3 Agent Market Arena results For EdgeQuant Agent over BTC window: 2026-05-11 to 2026-07-04. TSLA window: 2026-05-11 to 2026-06-26 (post-06-26 dates treated as flat for all agents due to an organiser-side payload issue).

Asset	Rank	Return	vs. B&H	Sharpe	Win Rate	Days	Final Balance
TSLA	**2nd**	+10.51%	+25.34%	3.34	67%	33	$110,335.34
BTC	**8th**	−0.52%	+22.41%	0.03	53%	55	$99,316.51

EdgeQuant Agent ranks 2^{nd} of 18 agents on TSLA, with a Cumulative Return of $+10.51\%$ (no-fee return: $+11.65\%$) over 33 active trading days (14 closed days) and a 67% win rate, outperforming buy-and-hold by $+25.34\%$. On BTC, the agent ranks 8^{th} of 18 agents, with a slightly negative fee-inclusive Cumulative Return of -0.52% (no-fee return: $+1.22\%$) over 55 active trading days and a 53% win rate, while still outperforming buy-and-hold by $+22.41\%$; this reflects a substantial buy-and-hold decline in BTC over the window, with the agent's balance remaining close to the $100,000 starting capital while the baseline fell markedly.

Early in the reporting window, the conviction-forcing SELL mandate identified a sustained TSLA price decline (from $445.08 on 2026-05-11 to $409.99 on 2026-05-18), maintaining short positions for six consecutive days (2026-05-13–2026-05-20); similarly, the BTC agent held a sustained short position from 2026-05-12 through 2026-05-19 during BTC's decline from $82,015 to $76,834. These early directional calls contributed to, but do not fully explain, the cumulative outcomes over the complete windows (33 trading days for TSLA, 55 for BTC).

Unlike the shorter reporting window used prior to the freeze, the current spans several weeks per asset, and win rates of 67% and 53% reflect a mix of correct and incorrect daily decisions rather than a single directional bet. The window remains shorter than our $\approx$270-day backtest period, however, and should still be read as an early field validation rather than a long-run track record.

6. Analysis and Discussion

The experimental results reveal a strong dependence between retrieval quality, market structure, and model-scale financial reasoning capacity. EdgeQuant Agent achieves its strongest performance on TSLA, where the pre-summarised news digests exhibit high correlation with short-term directional price movements. The combination of the skeptical analyst persona, catalyst-tier reasoning framework, and reflection-driven retrieval enables the system to consistently identify operational catalysts such as delivery variance, FSD monetisation signals, and competitive EV pressure. In contrast, BTC performance exposes an important architectural limitation of conviction-forcing trading systems. The mandatory BUY/SELL constraint frequently produces long exposure during sustained bearish regimes because institutional ETF-flow narratives remain semantically positive even while market structure

trends downward. This creates a mismatch between sentiment-driven retrieval and directional price momentum, indicating that conditional HOLD permission during extended downtrends is necessary for robust cryptocurrency trading.

The experiments further demonstrate that architectural improvements alone are insufficient without strong underlying reasoning models. The substantial Sharpe Ratio gap between GPT-OSS-120B and smaller models confirms that financial reasoning quality remains heavily dependent on model capability, particularly for multi-context retrieval synthesis and risk-sensitive decision making. Several additional limitations remain. First, the reported gains are attributed to the full four-tier memory architecture as a unit; due to time constraints, we were unable to conduct a per-tier ablation (e.g., removing the reflection layer or the long-term 10-K layer) to isolate the marginal contribution of each tier. Consequently, the relative importance of individual memory layers to the observed Sharpe Ratio improvement remains an open empirical question, and we flag this as a limitation of the present evaluation rather than a settled architectural claim. Second, the 3-day momentum signal is highly noisy for volatile assets with large intraday swings, occasionally amplifying short-term directional errors. Finally, session-level inconsistencies in memory ID assignment can create duplicate vector entries during checkpoint restoration, potentially affecting retrieval consistency across long evaluation windows. It is important to note that the performance of our EdgeQuant Agent, as well as the other agents listed on the leaderboard, is not consistent across BTC and TSLA. Our agent achieves its strongest performance on TSLA, where it ranks second of 18 agents on the leaderboard, while its performance on BTC is comparatively weaker (8th of 18), posting a small net-of-fees loss despite outperforming buy-and-hold by a wide margin. This discrepancy suggests that a single investment strategy may not generalize effectively across heterogeneous asset classes, as different markets exhibit distinct volatility structures, behavioral patterns, and underlying drivers.

7. Conclusion

We presented EdgeQuant Agent, an autonomous LLM-driven trading system for FinMMEval 2026 Task 3. The system's core contribution is a **four-tier cognitive memory architecture**—comprising short-term news, mid-term quarterly filings, long-term fundamental regime, and self-reflective decision memories—backed by a ChromaDB vector store with compound-scored semantic retrieval. A conviction-forcing mandate system explicitly combats the HOLD bias. With GPT-OSS-120B, the agent achieves a portfolio Sharpe Ratio of 3.165 ($+32.6\%$ over the equal-weight benchmark) with reduced maximum drawdown (8.22% vs. 11.75%), driven predominantly by the exceptional performance for TSLA (SR $= 5.152$, CR $= 40.61\%$).

On the official leaderboard, the agent secures **2nd place** (of 18 agents) on TSLA and **8th place** (of 18 agents) on BTC, outperforming buy-and-hold by $+25.34\%$ and $+22.41\%$ respectively.

Future Work. Several high-priority directions remain: (i) A conditional HOLD permission for BTC during periods including several days with negative momentum should be included into the model; (ii) an ablation study isolating each memory tier's contribution needs to be carried out; (iii) enhanced technical signals should be added (RSI, MACD, 20-day momentum); (iv) a post-trade memory importance reinforcement should be evaluated; (v) a multi-model ensemble using a majority vote could increase the robustness of EdgeQuant; and (vi) a full integration of the CVXPY optimiser output into the fractional position sizing should be implemented and tested.

Acknowledgments

The authors thank the FinMMEval 2026 organisers for providing the competition infrastructure, the `MBZUAI/finmmevallabclef2026` dataset, and the Agent Market Arena evaluation platform. We also acknowledge the open-source contributors of ChromaDB, CVXPY, FastAPI, sentence-transformers, and the Hugging Face ecosystem.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) for grammar checking and drafting assistance, and Google Gemini AI for generating and refining system architecture figures. After using these tools, the author(s) reviewed, edited, and validated all content and take full responsibility for the publication's content.

References

[1] A. Lopez-Lira, Y. Tang, Can ChatGPT forecast stock price movements? return predictability and large language models, arXiv preprint arXiv:2304.07619 (2023). URL: https://arxiv.org/abs/2304.07619.

[2] G. Koa, Y. Ma, Z. Wen, Q. Zheng, Can large language models beat wall street? unveiling the potential of AI in stock selection, in: Workshop on Generative AI for Finance (GenAI4Finance) at NeurIPS 2023, 2023. URL: https://arxiv.org/abs/2401.03737.

[3] S. Ouyang, H. Yun, X. Zheng, Ai as decision-maker: ethics and risk preferences of llms, 2025. URL: https://arxiv.org/abs/2406.01168.

[4] L. Chen, S. Liu, J. Yan, X. Wang, H. Liu, C. Li, K. Jiao, J. Ying, Y. V. Liu, Q. Yang, X. Li, Advancing financial engineering with foundation models: Progress, applications, and challenges, Engineering (Beijing) (2025). doi:10.1016/j.eng.2025.11.029.

[5] 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.

[6] 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.

[7] 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. URL: https://arxiv.org/abs/2510.11695. doi:10.48550/arXiv.2510.11695. arXiv:2510.11695.

[8] J. Trammer, Chroma Team, Chroma: The AI-native open-source embedding database, https://github.com/chroma-core/chroma, 2023. Accessed: 2026.

[9] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, D. Trajanov, A survey on sentiment analysis in finance: Experimentations and comparison of different methods with recent advances, IEEE Access 8 (2020) 131046–131086. doi:10.1109/ACCESS.2020.3009626.

[10] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, G. Mann, BloombergGPT: A large language model for finance, arXiv preprint arXiv:2303.17564 (2023). URL: https://arxiv.org/abs/2303.17564.

[11] H. Yang, X.-Y. Liu, C. D. Wang, FinGPT: Open-source financial large language models, in: FinLLM Symposium at IJCAI 2023, 2023. URL: https://arxiv.org/abs/2306.06031.

[12] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, J. Huang, FinMA: A financial large language model with pre-training and fine-tuning, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024. URL: https://arxiv.org/abs/2306.05443.

[13] Y. Yu, H. Li, Z. Chen, Y. Jiang, Y. Li, D. Zhang, R. Liu, J. W. Suchow, K. Khashanah, FinMem: A performance-enhanced LLM trading agent with layered memory and character design, arXiv preprint arXiv:2311.13743 (2024). URL: https://arxiv.org/abs/2311.13743.

[14] W. Zhang, L. Zhao, H. Xu, S. Shi, J. Sun, M. Li, X. Wang, R. Xia, X. Feng, Z. Zhao, Y. Li, Z. Zhang, Z. Zhao, C. Zhu, W. Zheng, J. Bian, FinAgent: A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist, arXiv preprint arXiv:2402.18485 (2024). URL: https://arxiv.org/abs/2402.18485.

[15] B. Zhao, K. Guo, R. Shi, J. Wen, F. Shi, Z. Zheng, H. Li, TRADER: An intelligent agent framework for automated trading with large language models, arXiv preprint arXiv:2411.09474 (2024). URL: https://arxiv.org/abs/2411.09474.

[16] P. Lewis, E. Perez, A. Piktus, F. Petra, V. Karpukhin, N. Ogunyemi, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474. URL: https://arxiv.org/abs/2005.11401.

[17] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), ACM, 2023. URL: https://arxiv.org/abs/2304.03442.

[18] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: Language agents with verbal reinforcement learning, in: Advances in Neural Information Processing Systems, volume 36, 2023. URL: https://arxiv.org/abs/2303.11366.

[19] J. He, M. Xu, Y. Hu, X. Yu, Y. Li, J. Wang, L. Guo, InvestorBench: A benchmark for financial question-answering in LLM-based investor agents, arXiv preprint arXiv:2412.18174 (2024). URL: https://arxiv.org/abs/2412.18174.

[20] M. Chen, A. Borole, A. Pal, From GPT-4 to gemini and beyond: Assessing the landscape of LLMs on the FLARE financial benchmarks, arXiv preprint arXiv:2403.11782 (2024). URL: https://arxiv.org/abs/2403.11782.

[21] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-Pack: Packaged resources to advance general chinese embedding, arXiv preprint arXiv:2309.07597 (2023). URL: https://arxiv.org/abs/2309.07597.

A. System Configuration

Table 6 summarises the complete hyperparameter configuration.

Table 6
EdgeQuant Agent complete configuration.

Parameter	Value
Trading symbols	BTC, TSLA
LLM temperature	0.2
Max new tokens	2048
Request timeout	300 s
Memory retrieval top-k	5 per layer
Momentum window	3 days
History window	10 days
Initial capital	$100,000
Portfolio lookback	3 days
Embedding model	`BAAI/bge-small-en-v1.5`
Embedding dimension	384
Shrinkage intensity β	0.1
Short decay α	0.90
Mid decay α	0.95
Long decay α	0.98
Reflection decay α	0.95
Memory importance cap	10.0
API port	7860

B. API Endpoint Specification

The competition endpoint accepts POST requests at `/trading_action/`:

```
# Sample request payload (TSLA)
{
  "date": "2026-05-15",
  "symbol": ["TSLA"],
  "price": {"TSLA": 175.0},
  "news": {"TSLA": ["Tesla␣Q1␣deliveries␣beat␣estimates..."]},
  "momentum": {"TSLA": "bullish"},
  "10k": {"TSLA": ["[SEC␣10-K␣Filing␣-␣2025-12-31]..."]},
  "10q": {"TSLA": ["[SEC␣10-Q␣Filing␣-␣2026-03-31]..."]},
  "history_price": {"TSLA": [
    {"date": "2026-05-12", "price": 173.50},
    {"date": "2026-05-13", "price": 174.20},
    {"date": "2026-05-14", "price": 175.00}
  ]}
}

# Response
{"recommended_action": "BUY"}
```

Valid actions are BUY, HOLD, and SELL. Any failure within the 3-minute timeout defaults to HOLD.

C. Prompt Templates

This appendix provides the complete prompt templates used by **EdgeQuant Agent** for both TSLA (equity) and BTC (crypto) assets across warmup and test phases.

C.1. TSLA Test-Phase Prompt

Investment information prefix:

```
"The current date is {cur_date} (Today: {day_of_week},
Tomorrow: {tomorrow_day}). Here are the observed financial market facts:
For {symbol}, the current price is ${cur_price:.2f}.
Historical Prices (last 10 days): {historical_prices}"
```

Research guidelines and tier classification (injected after short-term memory):

```
#### RESEARCH GUIDELINES:
1. CATALYST FRESHNESS: Primary Catalyst MUST be from
  'Short-term (TODAY'S CRITICAL NEWS)'.
2. CATALYST MAGNITUDE (M): Rate 1-5.
3. EXPECTATION VARIANCE (V): Positive/Negative Surprise?
4. BULL/BEAR TENSION: Pivot to highest Cash Flow impact.

#### TIER CLASSIFICATION:
- TIER 1 (Structural): Delivery misses/beats, Margin pivots.
- TIER 2 (Competitive): Market share shifts, Pricing wars.
- TIER 3 (Tactical): SEC noise, secondary macro signals.
```

Trading mandate (calendar-adaptive):

```
# Weekday:
"STRICT MANDATE: Weekday trading. You MUST choose BUY or SELL.
 HOLD is prohibited."

# Weekend:
"Note: Weekend trading (Saturday/Sunday). HOLD is permitted."
```

Final output instruction:

```
### INSTITUTIONAL RESEARCH COMMITTEE (TEST MODE):
State your ALPHA-driven directional stance for {symbol}.

### INVESTMENT CASE:
1. PRIMARY DRIVER: Isolate catalyst with highest TIER rating.
2. MAGNITUDE CHECK: Rate catalyst impact (1-5).
3. VARIANCE CHECK: Does it represent a trend-reversal surprise?
4. EXECUTION: Select decisive action. Avoid 'waiting' unless
  catalysts are flat (Magnitude < 2).

### COMPLIANCE:
- LIQUIDITY: Full fill at session close.
- CONVICTION: Managing a $100M book. Capture alpha aggressively.

### OUTPUT FORMAT (JSON):
{
 "investment_decision": "{allowed_decisions}",
 "reasoning": "[TIER X] Magnitude: [X/5] |
  Variance: [Surprise vs Trend] |
  Catalyst: [One-sentence impact rule] |
  Conviction: [High/Medium]."
}
```

C.2. BTC Test-Phase Prompt

Investment information prefix:

```
"The current date is {cur_date} (Today: {day_of_week},
Tomorrow: {tomorrow_day}). Here are the observed financial market facts:
For {symbol}, the current price is ${cur_price:.2f}.
Historical Prices (last 10 days): {historical_prices}"
```

Research guidelines (injected after short-term memory):

```
"Institutional crypto analysis focuses on 'Trust Variance' and
'Liquidity Flows'. Prioritize ETF net-inflows, corporate treasury
allocations (custody), and network-level security resilience.
Filter out retail-driven social media hype."
```

Trading mandate (unconditional all days):

```
"STRICT MANDATE: BTC always requires a high-conviction BUY or
 SELL decision. HOLD is prohibited."
```

Final output instruction:

```
### CRYPTO LIQUIDITY COMMITTEE (BITCOIN):
Identify 'Forceful Accumulation' or 'Structural Capitulation'.

### ANALYSIS STEPS:
1. TRUST VARIANCE: Strengthen or weaken 'Safe Haven' mandate?
2. LIQUIDITY FLOWS: ETF/Treasury movements (Big Money signal).
3. MAGNITUDE: Structural (ETF Inflow) or retail noise?
4. DIRECTIONAL BIAS: Eliminate HOLD bias.

### COMPLIANCE:
- DECISIVENESS: You MUST capture alpha. Passive HOLDING is
  for retail users, not professionals.

### OUTPUT FORMAT (JSON):
{
  "investment_decision": "{allowed_decisions}",
  "reasoning": "[Catalyst Type] |
    Net Flow Impact: [Positive/Negative] |
    Magnitude: [X/5] |
    Trust Score: [Current Delta]."
}
```