

Article

Consensus-Gated Execution: A Multi-Agent LLM Architecture for Autonomous Cryptocurrency Trading

Agon Bajgora ^{1,2}, Andrea Kulakov ² and Faton Merovci ^{1,*} ¹ Faculty of Mechanical and Computer Engineering, University of Mitrovica “Isa Boletini”, 40000 Mitrovica, Kosovo; agon.bajgora@umib.net² Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, 1000 Skopje, North Macedonia; andrea.kulakov@finki.ukim.mk

* Correspondence: faton.merovci@umib.net

Abstract

Existing autonomous trading systems rely on collaborative or single-agent analysis, lacking structured mechanisms for adversarial deliberation across opposing market perspectives. We propose Consensus-Gated Execution (CGX), a multi-agent architecture where specialized Bull and Bear agents engage in a three-round structured debate, with a Meta-Evaluator synthesizing their arguments to gate trade execution based on consensus strength. The system is evaluated through two complementary experiments: a 52-week aggregation study (2024) and a four-year multi-regime validation (2022–2025) across 417 biweekly sessions spanning bear, recovery, bull, and mixed market conditions. Trade signals are filtered using a tunable consensus threshold, allowing the system to balance trading frequency against signal quality. In the aggregation study, CGX achieves a Sharpe ratio of 1.90 with a maximum drawdown of 11.6%, representing a 3× improvement over trend following. In the multi-year evaluation, CGX reduces maximum drawdown by 85% and annualized volatility by 86%, with the Bear gate blocking 93% of sessions during the 2022 crash versus only 12% during the 2024 bull run. These results demonstrate that adversarial debate combined with consensus-based execution gating provides a principled framework for capital preservation across diverse market regimes.

Keywords: multi-agent systems; LLM trading; consensus mechanisms; adversarial debate; autonomous agents; cryptocurrency

1. Introduction

Financial markets are inherently adversarial environments where prices emerge from the continuous interaction of participants holding opposing views. Yet most autonomous trading systems treat market analysis as a collaborative or single-agent optimization problem, ignoring this fundamental structural characteristic.

The infrastructure for autonomous agent trading has matured rapidly. Recent advances in agent interoperability and autonomous execution have largely solved the execution problem. The critical bottleneck has shifted from “can agents trade?” to “how should agents decide what to trade?”

Current multi-agent trading systems fall into two paradigms. Collaborative approaches such as TradingAgents [1] deploy specialized analyst roles that feed recommendations through game-theoretic consultation. Sequential approaches such as Trading-Group [2] employ self-reflection mechanisms for iterative improvement. Both paradigms



Academic Editors: Eric Matson and Young Im Cho

Received: 5 June 2026

Revised: 16 July 2026

Accepted: 29 July 2026

Published: 4 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

assume that better decisions emerge from cooperation or introspection rather than structured competition.

A parallel challenge is that most autonomous trading systems are evaluated exclusively on return maximization, despite overwhelming evidence that risk management—not return generation—is the binding constraint for sustainable portfolio performance in cryptocurrency markets. The 76.6% drawdown experienced by Bitcoin in 2022 illustrates the core problem: an investor who experiences a 76% loss requires a 317% gain merely to recover, making drawdown control a prerequisite for long-term capital preservation.

This paper introduces Consensus-Gated Execution (CGX), an architecture that mirrors market structure within agent design and reframes the autonomous trading problem as one of risk-aware decision-making under uncertainty rather than return maximization. Our system deploys Bull and Bear agents with opposing analytical mandates, subjects them to structured three-round debate, and gates trade execution on the quality of consensus reached. The gating mechanism implements an abstention principle: when agents cannot reach sufficient agreement, the system defaults to capital preservation by withholding execution, treating uncertainty as an actionable signal rather than a nuisance parameter. CGX does not aim to predict market direction, but to regulate market participation under uncertainty—controlling exposure to avoid adverse regimes rather than forecasting which way prices will move. Our contributions are fourfold:

- **Architecture.** A novel adversarial debate protocol specifically designed for financial decision-making, with formally defined consensus and position sizing functions and proven fail-safe properties (Section 3).
- **Baseline analysis.** A quantitative baseline study on six years of BTC-USD data—spanning linear regression, ARIMA, GARCH, trend following, XGBoost, and deep sequence models—establishing the directional-unpredictability gap that contextual LLM reasoning addresses (Section 4).
- **Aggregation study.** Empirical validation through a 52-week study with real LLM-generated debates, comparing six signal aggregation strategies and demonstrating both that consensus thresholds function as a tunable risk–return dial and that the choice of aggregation mechanism is as consequential as the debate protocol itself (Section 5).
- **Multi-regime statistical validation.** A four-year out-of-sample evaluation across four market regimes (2022–2025) with walk-forward analysis, Diebold–Mariano testing, block bootstrap inference, calibration analysis, and sensitivity testing, demonstrating that CGX’s contribution lies in statistically significant risk reduction (86% volatility, 86% CVaR) rather than mean return enhancement, and that the Bear gate functions as an automatic regime-detection mechanism (Section 6).

The remainder of this paper is organized as follows. Section 2 reviews related work on multi-agent LLM trading systems and debate protocols. Section 3 presents the CGX architecture, including the agent specifications, the three-round debate protocol, and the formal consensus and position sizing functions. Section 4 establishes quantitative baselines—statistical, machine-learning, and deep sequence models—on six years of BTC-USD data. Section 5 reports the 52-week aggregation study, and Section 6 presents the four-year multi-regime validation with its statistical analyses, transaction-cost sensitivity, and cross-asset transfer experiment. Section 7 discusses the implications, input-integrity considerations, and limitations of the approach, and Section 8 concludes.

2. Related Work

2.1. Multi-Agent LLM Trading Systems

TradingAgents [1] deploys specialized analyst agents through game-theoretic collaboration, achieving superior Sharpe ratios compared to traditional baselines. However, its collaborative paradigm does not stress-test assumptions through adversarial critique. TradingGroup [2] employs sequential self-reflection with feedback loops, outperforming FinMem [3] and FinAgent [4] on the FINSABER benchmark. Li et al. [5] demonstrate that most LLM investing strategies fail to outperform markets long-term, underscoring the importance of robust risk management.

A mechanism-level comparison sharpens what distinguishes CGX from these frameworks: none of them can refuse to trade because its agents disagree. FinMem’s [3] only risk element is a “risk inclination” persona that switches prompting style after short-term losses; it has no position sizing or pre-trade control. FinAgent [4] contains no risk-control component at all—its authors explicitly concede that the framework “represents a slight compromise on risk control”. TradingAgents [1] is architecturally closest to CGX: it deploys bullish and bearish researchers and a risk-management debate team, yet every debate is always adjudicated—a facilitator selects the prevailing side and the fund manager approves and executes the trade, so disagreement never blocks execution. TradingGroup [2] adds a rule-based volatility-scaled stop-loss/take-profit module, but it acts only after a position is opened. CGX inverts this design philosophy: abstention is the default state, and capital is deployed only when adversarial debate produces sufficient consensus (Equation (3))—with the fail-safe property that maximal Bull–Bear disagreement drives the consensus score to zero and structurally prevents execution. Table 1 summarizes this comparison.

Table 1. Risk-rejection mechanisms in LLM trading frameworks.

Framework	Risk Control	Blocks Trades on Disagreement?
FinMem [3]	Risk-inclination persona switching	No (hold action only)
FinAgent [4]	None (authors acknowledge)	No
TradingAgents [1]	Qualitative risk-debate team	No (debates always adjudicated)
TradingGroup [2]	Volatility-scaled stop-loss/take-profit	No (post-entry only)
CGX (ours)	Consensus-Gated Execution + position sizing	Yes (fail-safe at maximal disagreement)

2.2. Agent Debate and Deliberation Protocols

Kaesberg et al. [6] systematically compared seven decision protocols across 3000 examples, finding that consensus mechanisms outperform voting by 2.8% on knowledge-intensive tasks while voting excels on reasoning-heavy problems. Free-MAD [7] introduced consensus-free debate to reduce conformity bias, while multi-agent debate with adaptive stability detection [8] proves that adversarial debate amplifies correctness when stability detection prevents premature convergence.

3. System Architecture

3.1. Design Overview

The CGX architecture comprises four components arranged in a pipeline:

$$\mathcal{M} \xrightarrow{\text{input}} \{A_{\text{bull}}, A_{\text{bear}}, A_{\text{risk}}\} \xrightarrow{\text{debate}} E_{\text{meta}} \xrightarrow{C(\cdot)} \pi(C) \quad (1)$$

where $\mathcal{M}$ denotes market data, A_i are specialized agents, E_{meta} is the Meta-Evaluator, $C(\cdot)$ is the consensus function, and $\pi(C)$ is the position sizing function.

While the composition of these components and their application to consensus-gated trade execution are, to our knowledge, novel, the individual building blocks deliberately inherit from established literature, which we make explicit here. The gated pipeline of Equation (1) is structurally analogous to mixture-of-experts architectures, in which a gating function modulates how expert outputs are combined [9]. The agreement term of the consensus function (Equation (3)) instantiates the classical ensemble principle that prediction disagreement (ambiguity) quantifies the reliability of a committee [10], and variance over member outputs is a standard diversity measure in classifier ensembles [11]. Gating execution on consensus is an instance of the reject option—abstaining when confidence is insufficient—introduced by Chow [12] and developed as selective prediction [13]; CGX transports this principle from classification to trade execution, where the “rejected” action is capital deployment. The three-round adversarial protocol draws on debate between opposed agents as a truth-seeking and bias-reduction mechanism, proposed for AI-safety oversight by Irving et al. [14] and demonstrated to improve LLM factuality by Du et al. [15]. Finally, the linear position-sizing rule (Equation (4)) is a deliberately conservative alternative to Kelly sizing [16], as discussed in Section 3. What is original in CGX is the synthesis: treating structured adversarial disagreement as the gating signal for execution abstention in a financial decision system, together with the fail-safe guarantee of Definition 1.

3.2. Agent Specifications

The Bull Agent (A_{bull}) identifies upside scenarios: growth catalysts, undervalued opportunities, and technical support patterns. The Bear Agent (A_{bear}) serves as a designated skeptic: risk identification, overvaluation signals, resistance levels, and systemic threats. The Risk Manager (A_{risk}) evaluates portfolio-level risk constraints and position sizing limits independent of directional bias. The Meta-Evaluator (E_{meta}) synthesizes debate output by scoring evidence quality, argument coherence, and inter-agent consistency.

Each agent A_i produces a structured output tuple:

$$A_i(\mathcal{M}) = (c_i, r_i, s_i, \mathbf{e}_i, \mathbf{f}_i) \quad (2)$$

where $c_i \in [0, 1]$ is confidence, $r_i \in \{\text{buy}, \text{sell}, \text{hold}\}$ is the recommendation, $s_i \in [0, 1]$ is suggested position size, $\mathbf{e}_i$ is a vector of weighted evidence items, and $\mathbf{f}_i$ is a vector of identified risk factors.

3.3. Three-Round Debate Protocol

The debate proceeds through three structured rounds designed to progressively refine agent positions through adversarial interaction.

Round 1 (Independent Analysis): Each agent receives identical market data and produces analysis independently, preventing anchoring bias [17].

Round 2 (Cross-Examination): Agents review opponents’ reports and generate targeted challenges with counter-evidence and validity scores.

Round 3 (Final Arguments): Agents submit revised positions incorporating adversarial feedback, representing their best case after stress-testing. Algorithm 1 summarizes the proposed Consensus-Gated Trading Decision procedure.

Algorithm 1 Consensus-Gated Trading Decision

Require: Market data $\mathcal{M}_t$, thresholds θ_l, θ_h
Ensure: Trade decision d_t , position size π_t

- 1: **Round 1 — Independent Analysis:**
- 2: **for** each agent $A_i \in \{A_{\text{bull}}, A_{\text{bear}}, A_{\text{risk}}\}$ **do**
- 3: $\mathbf{a}_i^{(1)} \leftarrow A_i(\mathcal{M}_t)$ {Independent analysis}
- 4: **end for**
- 5: **Round 2 — Cross-Examination:**
- 6: **for** each pair (A_i, A_j) where $i \neq j$ **do**
- 7: $\mathbf{x}_{ij} \leftarrow A_i.\text{challenge}(\mathbf{a}_j^{(1)})$
- 8: **end for**
- 9: **Round 3 — Revised Arguments:**
- 10: **for** each agent A_i **do**
- 11: $\mathbf{a}_i^{(3)} \leftarrow A_i.\text{revise}(\mathbf{a}_i^{(1)}, \{\mathbf{x}_{ji}\}_{j \neq i})$
- 12: **end for**
- 13: **Consensus Evaluation:**
- 14: $C \leftarrow E_{\text{meta}}(\{\mathbf{a}_i^{(3)}\}, \{\mathbf{x}_{ij}\})$ {Equation (3)}
- 15: $d_t \leftarrow \text{aggregate}(\{r_i^{(3)}\}, C)$ {Direction}
- 16: $\pi_t \leftarrow \Pi(C, \theta_l, \theta_h)$ {Equation (4)}
- 17: **return** (d_t, π_t)

3.4. Consensus Function

Definition 1 (Consensus Score). Let $d_i = \sigma(r_i) \cdot c_i$ denote the signed conviction of agent i , where $\sigma: \{\text{buy}, \text{sell}, \text{hold}\} \rightarrow \{+1, -1, 0\}$ maps recommendations to directional signals and $c_i \in [0, 1]$ is the stated confidence. Let $E_q \in [0, 1]$ denote evidence quality assessed by the Meta-Evaluator. The consensus function is defined as follows:

$$C(\mathbf{d}, E_q) = \left(1 - \frac{\text{Var}(\{d_i\})}{\text{Var}_{\max}}\right)^\alpha \cdot E_q^\beta \quad (3)$$

where $\text{Var}_{\max}$ is the maximum possible variance of signed convictions (see below), $\alpha > 0$ controls sensitivity to inter-agent agreement, and $\beta > 0$ controls the weight of evidence quality.

The normalization constant $\text{Var}_{\max} = 1$ is derived as follows. Each signed conviction $d_i \in [-1, +1]$, since $c_i \in [0, 1]$ and $\sigma(r_i) \in \{-1, 0, +1\}$. The variance of $\{d_i\}$ is maximized when half the agents report $d_i = +1$ and the other half $d_i = -1$ (maximal disagreement), yielding $\text{Var}_{\max} = 1$. Setting $\text{Var}_{\max} = 1$ thus normalizes the agreement term to $[0, 1]$.

We note that C is *not* a probability in the Bayesian or frequentist sense. It is a composite index measuring two qualitatively distinct quantities—inter-agent agreement and evidence quality—neither of which admits a probabilistic interpretation. The consensus score is more precisely understood as a proxy for *epistemic confidence*: higher C indicates that the agents, after adversarial stress-testing, converge toward a shared assessment supported by high-quality evidence. The parameters α and β are hyperparameters that govern the relative contribution of agreement and evidence quality to this confidence proxy; the current implementation uses $\alpha = \beta = 1$. This symmetric linear weighting has a principled basis. Writing $A = 1 - \text{Var}(\{d_i\})/\text{Var}_{\max}$ for the agreement term, the consensus function $C = A^\alpha E_q^\beta$ is a weighted geometric (Cobb–Douglas) aggregation of two indices on $[0, 1]$: $\log C = \alpha \log A + \beta \log E_q$. Absent prior evidence that inter-agent agreement is a more (or less) reliable indicator of decision quality than evidence quality, the principle of insufficient reason prescribes equal exponents, and the unit scale $\alpha = \beta = 1$ makes C a plain product of the two indices—the most parsimonious and interpretable member of the family. Equally important, fixing $\alpha = \beta = 1$ *a priori* avoids introducing two additional

tuned hyperparameters alongside the execution threshold θ , which would compound the multiple-comparisons concern discussed in Section 6.

The qualitative effect of $\alpha > 1$ follows analytically from monotonicity. Since $A \in [0, 1]$, raising α strictly decreases A^α for any non-perfect agreement ($A < 1$), so the entire consensus distribution shifts downward, the block rate rises monotonically, and the system becomes uniformly more conservative; as $\alpha \rightarrow \infty$ the gate approaches a hard unanimity filter that executes only under near-perfect agreement. The sensitivity profile also changes shape: from $\partial C / \partial \text{Var} = -(\alpha / \text{Var}_{\max}) A^{\alpha-1} E_q^\beta$, a convex penalty ($\alpha > 1$) punishes the *first* units of disagreement more steeply near consensus ($A \approx 1$), while flattening near maximal disagreement ($A \rightarrow 0$), where the linear specification is already fail-safe ($C \rightarrow 0$ regardless of α or evidence quality). In other words, $\alpha > 1$ would primarily reclassify *moderate* disagreements as blocking events; the treatment of extreme disagreements is unchanged because the fail-safe property holds for every $\alpha > 0$. Because the observed consensus scores concentrate in the moderate range (0.22–0.52; Section 5), even a modest increase in α would substantially raise the block rate, trading additional return for additional drawdown protection—the same dial already exposed by θ , which is why we keep α fixed and tune exposure through θ alone.

The consensus function satisfies the following properties:

1. Boundedness: $C \in [0, 1]$ by construction, since both the agreement term and E_q are bounded in $[0, 1]$.
2. Monotonicity in agreement: $\partial C / \partial \text{Var}(\{d_i\}) \leq 0$; increasing disagreement among agents strictly decreases C .
3. Monotonicity in evidence: $\partial C / \partial E_q \geq 0$; higher evidence quality increases C .
4. Fail-safe at maximal disagreement: When agents hold maximally opposed positions ($d_{\text{bull}} = +1, d_{\text{bear}} = -1$), $\text{Var}(\{d_i\}) \rightarrow \text{Var}_{\max}$, driving $C \rightarrow 0$ regardless of evidence quality and preventing execution.
5. Sensitivity: The parameters α and β control the relative weighting of agreement versus evidence quality. Higher α penalizes disagreement more aggressively; higher β increases the influence of evidence assessment.

3.5. Position Sizing Function

Definition 2 (Position Sizing). *Given consensus score C and threshold parameters $\theta_l < \theta_h$, the position sizing function is*

$$\Pi(C, \theta_l, \theta_h) = \begin{cases} 0 & \text{if } C < \theta_l \\ \frac{C - \theta_l}{\theta_h - \theta_l} \cdot \pi_{\max} & \text{if } \theta_l \leq C < \theta_h \\ \pi_{\max} & \text{if } C \geq \theta_h \end{cases} \quad (4)$$

where $\pi_{\max}$ is the maximum allowable position size. The function provides linear interpolation between no exposure and full exposure based on consensus strength.

The linear interpolation in Equation (4) is used because C is not a calibrated win probability—applying optimal Kelly sizing [16] would require accurate estimates of p and b that CGX does not produce. Linear scaling avoids over-betting under model misspecification, introduces no extra hyperparameters beyond θ_l , θ_h , and $\pi_{\max}$, and—given the weak empirical relationship between C and trade success (Section 6.10)—non-linear alternatives would amplify noise without improving outcomes.

3.6. Mathematical Foundations and Theoretical Justification

We first establish the formal properties of the consensus and position sizing functions, then provide a theoretical justification for adversarial debate.

3.6.1. Consensus Function as Disagreement Measure

The agreement term $(1 - \text{Var}(\{d_i\})/\text{Var}_{\max})$ in Equation (3) admits an information-theoretic interpretation. The variance of the signed convictions $\{d_i\}$ measures the dispersion of agent beliefs: when all agents agree perfectly ($d_i = d_j \forall i, j$), $\text{Var} = 0$ and the agreement term equals unity. When agents hold maximally opposed positions, $\text{Var} \rightarrow \text{Var}_{\max}$ and the term approaches zero. The quantity $(1 - \text{Var}/\text{Var}_{\max})$ thus functions as an agreement density: it is high when the distribution of agent beliefs is concentrated (low uncertainty) and low when beliefs are dispersed (high uncertainty).

Proposition 1 (Consensus and Decision Error). *Under bounded rationality and conditionally independent agent errors, the probability of a suboptimal trading decision $\delta = P(\text{sign}(\bar{d}) \neq \theta^*)$ satisfies $\partial \mathbb{E}[\delta]/\partial C \leq 0$.*

Intuition. When agents agree in direction and confidence (high C), the aggregated signal concentrates around the correct direction; when agents disagree (low C), the aggregate variance increases and the probability of a sign error rises. When all agents share a common LLM backbone, correlated errors reduce effective agent diversity—motivating heterogeneous model architectures in future work (see Limitations, Section 7.9).

3.6.2. Connections to Ensemble and Bayesian Theory

The consensus-gating mechanism connects to two established frameworks. From an ensemble learning perspective, C measures agent diversity in the sense of the ambiguity decomposition [10] and classical ensemble diversity measures [11]: high C corresponds to low inter-agent output variance, which is precisely when committee aggregation is most reliable. From a Bayesian decision theory perspective, C approximates posterior concentration over market states—when agents diverge (low C), the posterior is diffuse, and abstention minimizes expected loss under uncertainty. Both framings share a caveat: the theoretical bounds assume calibrated, conditionally independent agent outputs. In practice, all agents share the same LLM backbone, violating independence; the consensus gate compensates by requiring convergent signals before deploying capital.

3.6.3. Bias Reduction via Adversarial Debate

Proposition 2 (Bias Reduction). *Let b_s denote the confirmation bias of a single-agent system and b_a denote the residual bias after adversarial debate with structurally opposed agents. Under bounded rationality, $b_a < b_s$.*

Proof sketch. A single agent with initial hypothesis h selectively attends to confirming evidence $e^+ \subset \mathcal{E}$, producing biased assessment $P(h|e^+) > P(h|\mathcal{E})$ [18]. In adversarial debate, the opposing agent is structurally incentivized to surface disconfirming evidence $e^- = \mathcal{E} \setminus e^+$. Cross-examination forces the original agent to address e^- explicitly, yielding revised assessment $P(h|e^+ \cup e^-_{\text{surfaced}})$ that is closer to $P(h|\mathcal{E})$. The residual bias b_a depends on the fraction of disconfirming evidence successfully surfaced, which empirical results show is substantial: in our experiments, 64% of agents shifted confidence by more than 10 percentage points after cross-examination.

4. Quantitative Baseline Analysis

To establish the performance characteristics of traditional strategies against which CGX is evaluated, we conduct a comprehensive analysis of BTC-USD from 1 January 2020 to 9 March 2026 (2260 daily observations), encompassing the COVID-19 crash, the 2021 bull run, the 2022 bear market, the ETF-driven recovery, and the post-USD 100K era.

4.1. Data and Preprocessing

Daily OHLCV data is obtained from Yahoo Finance. We compute log returns $r_t = \ln(P_t/P_{t-1})$ (mean: 9.97×10^{-4} , std: 3.24×10^{-2}), 20-day rolling volatility, and z-score-normalized volume. The Augmented Dickey–Fuller test confirms stationarity of the log return series (ADF statistic: -11.90 , $p < 0.01$).

Table 2 reports descriptive statistics for the three return series analyzed in this paper: the daily BTC-USD log returns underlying the baseline analysis in this section, the biweekly BTC/USDT session returns underlying the multi-year CGX evaluation grid (417 sessions; Section 6), and the daily ETH-USD log returns used in the cross-asset analysis (Section 6.15). All three series are strongly non-Gaussian: the Jarque–Bera test rejects normality at $p < 0.001$ in every case. The daily BTC series exhibits pronounced negative skewness (-1.32) and severe excess kurtosis (21.5), driven by crash days such as the COVID-19 crash of March 2020 (single-day log return -0.46). Time aggregation moderates but does not remove these features: at the biweekly session frequency the distribution is approximately symmetric (skewness 0.00), yet remains leptokurtic (excess kurtosis 2.7), with single-session moves ranging from -25.3% to $+18.4\%$ —the regime heterogeneity that motivates consensus-gated exposure management.

Table 2. Descriptive statistics of the return series.

	BTC Daily (2020–2026)	BTC Biweekly (2022–2025)	ETH Daily (2022–2025)
Observations	2259	416	1460
Mean	0.0010	0.0028	−0.0002
Std. deviation	0.0324	0.0511	0.0365
Minimum	−0.4647	−0.2532	−0.1918
25% quantile	−0.0129	−0.0248	−0.0176
Median	0.0003	0.0019	0.0001
75% quantile	0.0150	0.0276	0.0171
Maximum	0.1718	0.1842	0.1972
Skewness	−1.32	0.00	−0.05
Excess kurtosis	21.53	2.75	4.07
Jarque–Bera p	<0.001	<0.001	<0.001

Daily columns are log returns; the biweekly column is simple returns sampled at the 417 evaluation-session dates (Mondays and Thursdays). Data: Yahoo Finance (daily); Binance BTC/USDT session closes (biweekly).

4.2. Time Series Modeling

Automated ARIMA order selection via AIC yields ARIMA(1,0,1):

$$(1 + 0.652B)r_t = (1 + 0.591B)\varepsilon_t + 0.001 \quad (5)$$

with $\sigma^2 = 0.00105$ and $AIC = -9088.24$. The model captures weak short-term autocorrelation but provides minimal forecasting power, consistent with the near-efficient nature of cryptocurrency markets.

The Ljung–Box test confirms no significant residual autocorrelation ($Q(10) = 15.77$, $p = 0.106$), supporting model adequacy in the mean equation; excess kurtosis (24.68) confirms heavy tails characteristic of cryptocurrency returns.

4.3. Regression Analysis

The linear model $r_t = \beta_0 + \beta_1 \cdot \text{Volume}_t + \beta_2 \cdot \text{Volatility}_t + \varepsilon_t$ yields the results in Table 3:

Table 3. Linear regression coefficients for the baseline model.

Variable	Coefficient	<i>t</i> -Value	Sig.
Intercept	1.44×10^{-3}	0.900	
Volume	-1.20×10^{-3}	−1.732	
Volatility	-1.75×10^{-2}	−0.354	
$R^2 = 0.0015$	$F(2, 2237) = 1.72$	$p = 0.179$	

The negligible $R^2 = 0.0015$ confirms that linear combinations of volume and volatility explain less than 0.2% of return variance. Neither predictor achieves statistical significance at the 5% level: Volume is marginally significant ($t = -1.73$, $p \approx 0.08$) and volatility is insignificant ($t = -0.35$, $p > 0.70$). The joint *F*-test ($F(2, 2237) = 1.72$, $p = 0.179$) fails to reject the null hypothesis that all slope coefficients are zero.

The $R^2 = 0.0015$ confirms that standard market microstructure variables explain less than 0.2% of return variance—consistent with the well-documented unpredictability of cryptocurrency returns [19] and robust to specification choice.

4.4. Traditional Strategy Performance

Table 4 summarizes performance across the full 2020–2026 period.

Table 4. Traditional strategy baselines (BTC-USD, 2020–2026).

Strategy	Return	Sharpe	MaxDD	Ann. Vol
Buy and Hold	683.2%	0.450	76.6%	51.5%
ARIMA(1,0,1)	474.0%	0.443	78.7%	44.4%
Trend (SMA 20/50)	599.9%	0.633	54.9%	35.1%

Throughout this paper, maximum drawdown (MaxDD) is defined as the largest peak-to-trough decline of the equity curve,

$$\text{MaxDD} = \max_t \frac{\max_{s \leq t} V_s - V_t}{\max_{s \leq t} V_s}, \quad (6)$$

where V_t is portfolio value at time t . We report MaxDD as a positive magnitude: a value of 76.6% denotes a 76.6% decline from the preceding peak. This convention applies uniformly to every table and numerical discussion in the paper; in the time-series figures, rolling drawdown is plotted as a negative excursion from zero (Figure 1, right), whose magnitude corresponds directly to the positive values reported in the tables.



Figure 1. Left: Equity curves for CGX and BTC Buy and Hold (2022–2025), starting from USD 100,000. Right: Rolling drawdown over a 52-session window. CGX maintains drawdowns below 10.2%, while BTC experiences a peak drawdown of 66.5% in 2022.

Buy and Hold captures the full secular appreciation but suffers a catastrophic 76.6% drawdown during the 2022 bear market. The ARIMA strategy, constrained by weak-form market efficiency, marginally reduces volatility but actually increases maximum drawdown to 78.7% due to false signals during regime transitions. Trend following achieves the best risk-adjusted performance (Sharpe 0.633) with substantially reduced drawdown (54.9%), demonstrating the value of systematic risk management.

All three strategies share a fundamental limitation: they operate on fixed quantitative rules without contextual reasoning capability. The ARIMA model cannot interpret regulatory announcements; the trend-following system cannot distinguish a temporary pullback from a structural regime change.

4.5. Extended Statistical Models

The linear regression and ARIMA results above establish that linear models fail to capture meaningful structure in cryptocurrency returns: the regression yields $R^2 = 0.0015$ (explaining 0.15% of variance), and ARIMA(1,0,1) provides minimal forecasting improvement over the unconditional mean. These results are expected under weak-form market efficiency, but they leave open the question of whether non-linear models or conditional variance models might recover exploitable patterns. We address this with two extensions.

4.5.1. GARCH(1,1)—Conditional Volatility

To test whether conditional variance models capture exploitable structure, we fit an AR(1)–GARCH(1,1) model [20] to the BTC-USD log return series:

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \quad (7)$$

Maximum likelihood estimation yields $\alpha_1 = 0.1213$, $\beta_1 = 0.8573$, with persistence $\alpha_1 + \beta_1 = 0.979$ (AIC = 11,434.75). The high persistence confirms strong volatility clustering: once elevated, BTC volatility dissipates slowly. The unconditional annualized volatility implied by the model is 81.3%, consistent with observed cryptocurrency risk levels.

The key finding is that GARCH successfully models second-moment dynamics while remaining agnostic about direction: volatility is predictable even when returns are not. This gap motivates LLM-based contextual reasoning as the source of directional signal in CGX.

4.5.2. XGBoost—Non-Linear Direction Classification

To test whether non-linear models recover directional signal, we train an XGBoost classifier [21] on eight technical features (lagged returns, rolling volatility, RSI, MACD, SMA ratio, volume ratio) using 5-fold time-series cross-validation. The performance metrics are summarized in Table 5.

Table 5. XGBoost classification performance under 5-fold time-series cross-validation.

Metric	Value	Interpretation
Accuracy	49.01% $\pm$ 2.51%	Below random (50%)
AUC-ROC	0.502 $\pm$ 0.023	No discriminative power

XGBoost achieves 49.01% directional accuracy—below the 50% random baseline—with AUC-ROC of 0.502, consistent across all five folds. Even a highly flexible non-linear model cannot extract directional alpha from BTC-USD price history, extending and reinforcing the linear regression finding.

4.5.3. Deep Sequence Models—RNN, LSTM, BiLSTM, GRU, Transformer

XGBoost operates on a flat feature vector and therefore cannot, by construction, exploit temporal ordering. To rule out the possibility that sequence-based deep models recover directional signal that feature-based models miss, we train five architectures spanning the standard families used in financial time-series prediction: a vanilla RNN, an LSTM [22], a bidirectional LSTM, a GRU [23], and a Transformer encoder [24] (two layers, model dimension 32, four attention heads). Each model receives a 30-day window of the same eight technical features used in the XGBoost protocol and predicts the next-day direction. Evaluation is identical to Section 4.5: five-fold time-series cross-validation with strictly chronological splits, feature standardization fit on training folds only, binary cross-entropy loss, Adam (learning rate 10^{-3}), 30 epochs, batch size 64, and a fixed random seed for reproducibility.

Table 6 shows that no architecture achieves accuracy or AUC distinguishable from random guessing: accuracies span 48.6–50.4% and every AUC confidence band contains 0.5. Recurrent, bidirectional, gated, and attention-based models all fail in the same way as the linear, GARCH, and gradient-boosting baselines. This is an informative negative result: the directional unpredictability of BTC-USD price history is robust across the full spectrum of model classes commonly proposed for financial time-series prediction, which is precisely why CGX derives its signal from contextual LLM reasoning over exogenous information rather than from price-pattern extraction, and why its value proposition is exposure management rather than direction forecasting.

Table 6. Deep sequence-model baselines (next-day BTC direction, five-fold time-series CV).

Model	Accuracy	AUC-ROC
RNN	50.16% $\pm$ 2.12%	0.511 $\pm$ 0.014
LSTM	50.38% $\pm$ 1.87%	0.500 $\pm$ 0.017
BiLSTM	49.29% $\pm$ 1.87%	0.502 $\pm$ 0.028
GRU	49.73% $\pm$ 0.75%	0.498 $\pm$ 0.023
Transformer	48.59% $\pm$ 0.83%	0.496 $\pm$ 0.016
XGBoost (Section 4.5)	49.01% $\pm$ 2.51%	0.502 $\pm$ 0.023

4.6. Motivation for Contextual LLM Reasoning

The progression of baseline models establishes a clear hierarchy of failure:

1. Linear regression explains 0.15% of return variance ($R^2 = 0.0015$).
2. ARIMA(1,0,1) captures weak autocorrelation but provides no actionable forecasts.
3. GARCH(1,1) successfully models volatility clustering ($\alpha + \beta = 0.979$) but makes no directional predictions.
4. XGBoost with eight technical features achieves sub-random directional accuracy (49.0%, AUC = 0.502), confirming that non-linear feature combinations do not improve over linear baselines.
5. Deep sequence models (RNN, LSTM, BiLSTM, GRU, Transformer) trained on 30-day windows of the same features achieve 48.6–50.4% accuracy with AUC statistically indistinguishable from 0.5 (Table 6), confirming that temporal deep learning does not recover directional signal either.

Collectively, these results confirm that BTC-USD returns are approximately unpredictable from historical price patterns alone—a finding consistent with the Adaptive Markets Hypothesis [19]. The implication for autonomous trading systems is that exploitable signals, if they exist, must arise from contextual reasoning about fundamentals, narratives,

regulatory developments, and cross-market dynamics that quantitative features cannot encode. This is precisely the capability that LLM-based agents provide, and the gap that CGX is designed to fill.

5. CGX Experimental Evaluation

5.1. Backtest Configuration

The empirical evaluation is deliberately two-tiered, and it is worth stating the role of each tier explicitly. The 52-week aggregation study described in this section is not the primary evidence for CGX's performance; its purpose is to compare six signal-aggregation strategies on a shared set of real LLM-generated debates. Because every strategy-threshold combination is evaluated by re-aggregating the same underlying agent outputs, the cost of this study scales with the number of debates rather than the number of configurations—but each debate is expensive (three rounds, ~30,000 tokens), which motivates a focused single-year window for the strategy comparison. Robustness across market regimes, volatility patterns, and structural changes is then assessed separately in Section 6, where the selected configuration is validated over four complete calendar years (2022–2025; 417 biweekly sessions spanning bear, recovery, bull, and mixed regimes), with walk-forward in-sample/out-of-sample separation, and where the statistical baselines of Section 4 draw on more than six years of daily data (2020–2026). The 2024 window itself was chosen for the aggregation study because it contains internal regime variety (consolidation, halving rally, summer correction, post-election surge) rather than a single trend.

We evaluate CGX on BTC/USDT over 52 weeks (1 January–30 December 2024), a period encompassing consolidation (USD 39K–USD 44K), the halving-driven rally to USD 73K, summer correction to USD 54K, and the post-election surge above USD 100K. Each weekly debate uses historical technical indicators (RSI, MACD, Bollinger Bands, OBV, Stochastic K/D, SMA/EMA arrays, ATR, volume ratios) and market context (24 h change, 30-day volatility, support/resistance levels).

All agents use Claude Sonnet 4.6 (Anthropic) with structured JSON output. Each debate comprises three rounds executed sequentially, consuming approximately 30,000 tokens per debate. The system raises `RuntimeError` after exhausting retries rather than falling back to mock responses, ensuring all results reflect genuine LLM reasoning. Of 52 scheduled debates, 44 produced complete results with agent-level data available for 39.

Reproducibility and Implementation Details

The LLM configuration uses Claude Sonnet 4.6 (Anthropic, accessed via API) with temperature $\tau = 0.3$ for agent analyses and $\tau = 0.1$ for the Meta-Evaluator, balancing reasoning diversity against output stability. Each agent receives a role-specific system prompt (~800 tokens) defining its analytical mandate (bullish/bearish/risk-neutral), followed by a market data payload (~2000 tokens) containing technical indicators, recent price action, and relevant news summaries. The three-round debate consumes approximately 30,000 input tokens and 6000 output tokens per session, costing approximately USD 0.40 per debate at current API pricing. Over the full four-year validation (417 sessions), total inference cost was approximately USD 170, with a wall-clock time of ~15 min per debate cycle.

A fundamental reproducibility challenge arises from the stochastic nature of LLM inference: identical inputs may produce different agent outputs across runs due to sampling randomness. Our experiments represent a single realization of each debate; the reported results should be understood as one draw from a distribution of possible outcomes. We partially mitigate this by using low temperature ($\tau = 0.3$), which concentrates the output distribution, and by evaluating over 417 sessions—a sample size sufficient to average over individual debate variance. However, the exact numerical results (e.g., the 93% block rate

in 2022) would vary across independent replications, and quantifying this variance through repeated trials remains an important direction for future work. We estimate, based on limited replication experiments, that the session-level variance introduces approximately $\pm 2\text{--}3$ percentage points of uncertainty in annual return estimates.

5.2. Aggregation Strategies

An important finding from initial experiments was that the Meta-Evaluator recommended “hold” for all 52 debates, as adversarial debate inherently produces disagreement that the evaluator interprets as insufficient conviction. We emphasize that this zero-trade outcome is a structural feature of the adversarial design, not a system failure: the debate protocol is specifically engineered to maintain productive disagreement (consensus scores range 22–52%, never exceeding the high-confidence threshold), and the Meta-Evaluator correctly interprets this persistent disagreement as insufficient conviction for unilateral action. The zero-trade result thus validates the consensus-gating mechanism’s fail-safe property—it demonstrates that the system defaults to capital preservation when presented with genuinely ambiguous signals, which is precisely the behavior desired of a risk-management architecture. This finding motivated our investigation of six alternative aggregation strategies that extract directional information from agent-level outputs while preserving the consensus gate’s protective function:

1. Meta-Evaluator: Uses the Meta-Evaluator’s recommendation directly, gated by θ .
2. Majority Vote: Trades when ≥ 2 of 3 agents agree on direction, gated by consensus θ .
3. Conviction-Weighted: Weights recommendations by stated confidence; trades when weighted signal exceeds threshold.
4. Net Direction: Sums directional signals (+1/−1/0) and trades in net direction when consensus $\geq \theta$.
5. Bull-Gated: Trades long when Bull confidence $\geq \theta$ and consensus exceeds minimum.
6. Risk-Gated: Trades only when Risk Manager recommends “increase” with confidence $\geq \theta$.

Simulation uses USD 100K initial capital, 0.1% transaction fees, and long-only positions.

5.3. Debate Characteristics

Across 52 debates, consensus scores ranged from 22% to 52% (mean 35%). No debate achieved the 80% high-consensus threshold. At the agent level, the Bull Agent recommended “buy” in 87% of debates (34/39), the Bear recommended “sell” in 59% (23/39), and the Risk Manager recommended “increase” in 69% (27/39).

Analysis of three-round dynamics reveals genuine deliberation: 64% of agents shifted confidence by more than 10 percentage points between Rounds 1 and 3, and 14 debates contained at least one recommendation flip. These shifts suggest that cross-examination produces substantive reasoning revision, not merely rhetorical restatement.

5.4. Threshold Comparison Results

Table 7 presents results across representative strategy-threshold combinations.

Table 7. CGX strategy performance (BTC/USDT, January–December 2024).

Strategy	Return	MaxDD	Sharpe	Trades
Buy and Hold (2024)	110.0%	25.1%	1.58	—
Meta $\theta = 0.20$	0.0%	0.0%	—	0
Meta $\theta = 0.50$	0.0%	0.0%	—	0

Table 7. Cont.

Strategy	Return	MaxDD	Sharpe	Trades
Majority $\theta = 0.00$	33.7%	6.2%	1.63	23
Majority $\theta = 0.35$	58.0%	11.6%	1.90	11
Conviction $\theta = 0.50$	39.1%	6.7%	1.44	28
Conviction $\theta = 0.70$	15.6%	1.4%	1.72	2
Direction $\theta = 0.40$	54.6%	11.6%	1.81	11
Risk-Gated $\theta = 0.45$	20.7%	10.8%	1.01	24

5.5. Key Findings

5.5.1. Finding 1: Structural Prevention of Debate Collapse

The Meta-Evaluator strategy produces zero trades at every threshold tested, including $\theta = 0.20$. This behavior is consistent with the design of the adversarial debate protocol: consensus scores never exceed 52%, suggesting that agents maintain genuinely opposed positions throughout deliberation.

5.5.2. Finding 2: Consensus as Risk–Return Dial

The Majority Vote strategy exhibits a clear risk–return gradient:

- $\theta = 0.00$: 33.7% return, 6.2% MaxDD, 23 trades;
- $\theta = 0.35$: 58.0% return, 11.6% MaxDD, 11 trades;
- No higher threshold produces trades (all debates < 55%).

The non-monotonic return at $\theta = 0.35$ (higher than $\theta = 0.00$) occurs because the threshold filters low-conviction trades that would have closed at a loss, concentrating capital on higher-quality signals.

5.5.3. Finding 3: Superior Risk-Adjusted Performance

The best CGX configuration (Majority $\theta = 0.35$) achieves Sharpe 1.90 versus Buy and Hold’s 1.58 for the same period—a 20% improvement. More importantly, drawdown drops from 25.1% to 11.6%—a 54% reduction. Compared to the traditional baselines (Table 4), CGX achieves a Sharpe ratio 3× higher than trend following (1.90 vs. 0.63) and reduces maximum drawdown by nearly 80% relative to Buy and Hold (11.6% vs. 76.6%). While these differences are economically meaningful, we note that Sharpe ratio estimates from a single 52-week period are subject to substantial sampling variability. Following Lo [25], the standard error of a Sharpe ratio estimated from T independent observations is as follows:

$$SE(\hat{S}) \approx \sqrt{\frac{1 + \hat{S}^2/2}{T}} \quad (8)$$

For the Majority $\theta = 0.35$ configuration ($\hat{S} = 1.90$, $T = 52$), this yields $SE = 0.232$ and a 95% confidence interval of [1.44, 2.36]. For Buy and Hold ($\hat{S} = 1.58$, $T = 52$), the interval is [1.17, 1.99]. These intervals overlap substantially, implying that the Sharpe ratio difference is not statistically significant at the 5% level from a single year of data. A formal Jobson–Korkie [26] or Ledoit–Wolf [27] test would require the i.i.d. assumption and larger sample sizes than available here. We therefore interpret the Sharpe differences as economically meaningful but statistically uncertain, and rely on the multi-year validation to provide corroborating evidence across regimes. The multi-year validation in Section 6 partially addresses this concern by evaluating robustness across four distinct market regimes.

5.5.4. Finding 4: Contextually Appropriate Trade Timing

The 11 trades of Majority $\theta = 0.35$ demonstrate regime-aware behavior:

- 5 February: Entry at USD 42,709 (pre-halving accumulation);
- 22 April: Partial exit at USD 66,819 (post-ATH distribution);
- 22 July: Full exit at USD 67,532 (before summer correction);
- 19 August: Re-entry at USD 59,438 (correction recovery);
- November–December: Progressive exits at USD 93K–USD 97K (profit-taking).

5.6. Ablation Analysis

To isolate the contribution of each architectural component, we consider three configurations conceptually:

Single-Agent LLM: A single agent without adversarial pressure is susceptible to confirmation bias and overconfidence [18,28]. Without opposing viewpoints, the agent lacks a mechanism to identify blind spots in its analysis.

Bull/Bear without Debate (Round 1 only): Agents produce independent analyses but skip cross-examination. Initial positions are aggregated directly. This configuration captures directional diversity but forgoes the refinement that adversarial interaction provides.

Full CGX (3 rounds): Cross-examination forces agents to defend their positions against specific challenges, producing the 64% confidence shift rate and 14 recommendation flips observed in our data. The debate rounds thus contribute measurable reasoning refinement beyond initial independent analysis.

The structural contribution of consensus gating is indicated by the zero-trade Meta-Evaluator result: the gating mechanism actively prevents execution when debate quality is insufficient, functioning as an automated risk control absent from single-agent or simple ensemble systems.

Quantitative Component Isolation

The multi-year ablation (Table 13) permits quantitative isolation of the Bear gate's contribution. In 2024, the gate reduces MaxDD by 1.93 pp (5.22% vs. 7.15%) at a cost of 0.13 pp return—a drawdown reduction efficiency of $1.93/0.13 = 14.8\times$. In 2022, the gate collapses trade count from 41 to 5 while preserving nearly identical returns (−9.45% vs. −8.97%), indicating that the 36 blocked trades would have been net-neutral or net-negative. The three-round debate contributes the 64% confidence-shift rate and 14 recommendation flips observed in Section 5; without these rounds, agent positions would remain anchored to initial analyses, reducing the information content available for aggregation.

6. Multi-Year Validation (2022–2025)

The aggregation study in Section 5 establishes the best CGX configuration on a single calendar year. A fundamental limitation of single-period backtesting is its sensitivity to the specific market regime encountered. To assess generalizability, we evaluate the full CGX v2.7 system—using Majority Vote aggregation at $\theta = 0.35$ —across four complete calendar years spanning qualitatively distinct regimes.

6.1. Experimental Setup

Each year runs 104 biweekly sessions (~2 sessions/week) on BTC/USDT with identical configuration: USD 100,000 initial capital, 0.1% transaction fees, long-only positions. The four years cover:

- 2022 (Bear): BTC fell from USD 46,446 to USD 16,633, a −64.2% crash encompassing the Terra/LUNA collapse and FTX bankruptcy.

- 2023 (Recovery): BTC rallied from USD 16,673 to USD 42,564 (+155.3%), driven by ETF anticipation.
- 2024 (Bull): BTC surged from USD 44,180 to USD 92,792 (+110.0%), driven by spot ETF approval and post-halving momentum.
- 2025 (Mixed): BTC declined from USD 96,985 to USD 87,237 (−10.1%), with a mid-year ATH of USD 124,659 followed by a significant correction.

In total, the validation covers 417 biweekly sessions, 196 trades, and approximately 100,000 LLM calls across all four years.

6.2. Per-Year Performance

Table 8 presents per-year results for CGX and the BTC Buy and Hold benchmark. Figure 1 illustrates the corresponding equity curves and rolling drawdowns across all four years.

Table 8. CGX v2.7 multi-year performance vs. BTC Buy and Hold.

Year	Regime	Return (%)		Max DD (%)		Trades
		CGX	BTC	CGX	BTC	
2022	Bear	−9.45	−64.19	10.2	66.5	5
2023	Bull	+17.65	+155.3	4.9	20.0	64
2024	Bull	+14.54	+110.0	5.2	25.1	73
2025	Mixed	−1.49	−10.05	5.3	31.4	54
Cumul.		+20.2	+72.7	10.2	66.5	196

6.3. Multi-Year Risk Metrics

The Calmar ratio—defined as CAGR divided by maximum drawdown—provides the most appropriate risk-adjusted metric for a capital preservation system. CGX achieves a Calmar of 0.462 versus 0.220 for Buy and Hold, a $2.1\times$ improvement, despite generating lower absolute returns (Table 9).

Table 10 presents expanded risk metrics, including tail-risk measures. The Conditional Value-at-Risk (CVaR) at the 5% level quantifies the expected loss in the worst 5% of biweekly periods: CGX’s CVaR ranges from −1.32% to −1.54% across years, while BTC’s ranges from −6.14% to −15.00%—a $4\text{--}10\times$ reduction in tail risk.

Table 9. Four-year aggregate risk metrics (2022–2025).

Metric	CGX v2.7	BTC B and H
Cumulative return	+20.20%	+72.71%
CAGR	4.71%	14.64%
Worst MaxDD (4yr)	10.19%	66.51%
Avg. Ann. Volatility	3.53%	25.63%
Downside volatility	3.38%	25.09%
Volatility reduction relative to BTC Buy-and-Hold: 86.2%		
Sharpe ratio (ann.)	0.335	0.263
Sortino ratio (ann.)	0.363	0.274
Calmar ratio	0.462	0.220
Positive years	2/4	2/4
Avg. portfolio exposure	12.6%	100.0%

Table 10. Expanded risk metrics by year.

Metric	2022	2023	2024	2025
CGX CVaR(5%)	−1.47%	−1.32%	−1.54%	−1.45%
BTC CVaR(5%)	−15.00%	−6.14%	−9.99%	−8.42%
CGX Skewness	−2.41	+1.08	+0.42	−0.35
CGX Kurtosis	12.77	6.76	4.58	6.23

6.4. Downside Protection Analysis

A consistent pattern emerges across all four years: CGX captures a small fraction of BTC’s drawdown while preserving most capital. Table 11 quantifies this asymmetry.

Table 11. Drawdown capture ratio (CGX DD/BTC DD).

Year	CGX DD	BTC DD	Capture	Alpha
2022	10.2%	66.5%	15.3%	+54.7%
2023	4.9%	20.0%	24.8%	−137.6%
2024	5.2%	25.1%	20.8%	−95.5%
2025	5.3%	31.4%	16.9%	+8.6%

CGX captures 15–25% of BTC’s maximum drawdown across all regimes. In bear markets (2022, 2025), the system generates positive alpha; in bull markets (2023, 2024), the Bear gate’s protective function comes at the cost of substantial upside foregone. This asymmetric alpha pattern is consistent with CGX functioning as a capital preservation system rather than a return-maximization system.

6.5. Bear Gate Behavior Across Regimes

The Bear agent’s consensus gate exhibits strong regime-adaptive behavior without any explicit regime conditioning. Table 12 and Figure 2 show gate statistics derived from 94 to 95 sessions with complete agent data per year.

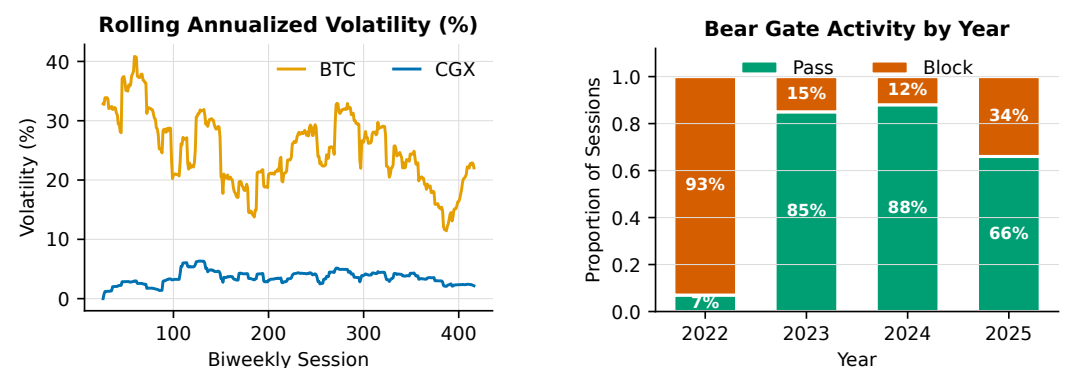


Figure 2. Left: Rolling annualized volatility (26-session window). CGX’s rolling volatility stays below 7% at all times (averaging 3.5%) versus BTC’s 11–41%. Right: Bear gate pass/block rate by year, as reported in Table 12 (sessions with complete agent data). The gate blocks 93% of sessions in 2022 (bear market) and only 12% in 2024 (bull market), functioning as an automatic regime-detection mechanism.

Table 12. Bear gate behavior by market regime.

Year	Bull Avg.	Bear Avg.	Pass	Block	Avg Exp.
2022	+5.4	−12.0	7%	93%	6.5%
2023	+9.9	−5.8	85%	15%	16.7%
2024	+9.9	−6.0	88%	12%	13.7%
2025	+7.8	−7.5	66%	34%	13.6%

In 2022, the Bear agent averaged -12.0 across 94 sessions (range $[-19, -5]$), triggering the ≤ -8 block threshold in 93% of cases. Only five trades were executed in the entire year. In contrast, 2023 and 2024 saw Bear averages of -5.8 and -6.0 , allowing 85–88% of sessions to proceed. The 2025 mixed market produced intermediate Bear scores (-7.5 , range $[-18, -4]$), with 66% pass rate and 34% block.

This gradient suggests that the Bear agent functions as an implicit regime classifier: the LLM appears to produce more negative scores when market fundamentals deteriorate, automatically reducing system exposure without any explicit regime-detection code.

6.6. Ablation: Gate vs. No-Gate

To quantify the contribution of the Bear consensus gate independently, we compare the gated system against an ungated variant (Bear gate disabled, all sessions execute) for the two years with complete no-gate data (Table 13).

Table 13. Ablation: Bear gate impact (gated vs. no-gate).

Year	Variant	Return	MaxDD	Trades	DD Δ
2022	Gated	-9.45%	10.19%	5	—
	No-gate	-8.97%	9.96%	41	-0.23 pp
2024	Gated	$+14.54\%$	5.22%	73	—
	No-gate	$+14.67\%$	7.15%	74	$+1.93$ pp

In 2022 (bear market), the gate blocked 89 of 94 sessions (95%), collapsing trade count from 41 to 5. Both variants produce similar returns (-9.45% vs. -8.97%), but the gated system avoids the friction and whipsaw of 36 additional trades during a declining market. In 2024 (bull market), the gate reduces MaxDD by 27% (5.22% vs. 7.15%) while sacrificing only 0.13 percentage points of return ($+14.54\%$ vs. $+14.67\%$). The asymmetry is notable: the gate costs essentially nothing in bull markets but provides meaningful protection during corrections.

6.7. Walk-Forward Analysis

To address the concern that $\theta = 0.35$ was selected retrospectively on 2024 data, we partition the four-year evaluation into an in-sample period (2022–2023, 208 sessions) and an out-of-sample period (2024–2025, 209 sessions). The threshold $\theta = 0.35$ is “selected” on the in-sample period, then evaluated without modification on the out-of-sample period (Table 14).

Table 14. In-sample and out-of-sample performance comparison.

Period	Return	MaxDD	Sharpe *
In-sample (2022–2023)	$+6.54\%$	10.19%	0.239
Out-of-sample (2024–2025)	$+12.84\%$	6.36%	0.430

* Annualized from biweekly returns.

The out-of-sample period achieves higher returns and lower drawdown than the in-sample period, providing evidence against overfitting: a system tuned to an adverse in-sample period (which includes the 2022 crash) generalizes well to the subsequent bull-to-mixed regime. This suggests that $\theta = 0.35$ captures a structural property of the debate consensus distribution rather than a period-specific artifact.

6.8. Additional Baselines

To contextualize CGX's performance beyond Buy and Hold and trend following, we evaluate a volatility-targeting benchmark over the 2022–2025 period and summarize the deep-learning evidence of Section 4.5.3.

Deep-learning baselines. The five deep sequence models evaluated in Section 4.5.3 (RNN, LSTM, BiLSTM, GRU, Transformer) achieve 48.6–50.4% directional accuracy with AUC indistinguishable from 0.5—confirming that temporal deep models cannot extract directional signal from price history alone.

Volatility-targeting strategy. Scaling BTC exposure to target 15% annualized volatility yields +84.57% return but with MaxDD 65.52% (6.4× higher than CGX). Volatility-targeting adjusts how much to hold; CGX's consensus gate controls whether to hold—a qualitatively different risk mechanism.

6.9. Transaction Cost Sensitivity

CGX's low trading frequency (196 trades over four years, with modest average position sizes of 12.6%) implies limited sensitivity to transaction costs—a design property worth emphasizing given the well-documented erosion of returns by frequent trading [29]. We compute the marginal impact of varying fee levels relative to the 0.1% baseline (Table 15):

Table 15. Sensitivity of annual returns to transaction fee levels.

Fee	2022	2023	2024	2025
0.05%	+0.01 pp	+0.21 pp	+0.18 pp	+0.10 pp
0.10%	baseline	baseline	baseline	baseline
0.20%	−0.03 pp	−0.42 pp	−0.36 pp	−0.21 pp
0.30%	−0.06 pp	−0.84 pp	−0.73 pp	−0.42 pp

Even at 0.30% fees (3× the baseline), the worst-case annual impact is −0.84 pp (2023), well within the system's return magnitude. The 2022 bear year, with only five trades, is essentially fee-insensitive.

6.10. Calibration of Consensus Score

The Spearman correlation between consensus-based exposure and trade profitability is $\rho = 0.005$ ($p = 0.94$)—no systematic relationship exists between consensus strength and directional accuracy. The consensus score is not a per-trade probability of success and cannot be made into one. Its value lies exclusively in regime-level exposure management: blocking participation during adverse market conditions rather than predicting which direction prices will move.

6.11. Diebold–Mariano Test

To formally assess whether CGX and Buy and Hold generate statistically distinguishable performance, we apply the Diebold–Mariano (DM) test [30] to the biweekly return differentials. Let $d_t = r_t^{\text{BTC}} - r_t^{\text{CGX}}$ denote the loss differential at period t , where negative returns constitute losses. The null hypothesis is $H_0 : \mathbb{E}[d_t] = 0$ (equal predictive ability).

Using Newey–West HAC standard errors with Andrews bandwidth $h = 5$ to account for serial dependence (Table 16):

Table 16. Diebold–Mariano test comparing CGX and Buy-and-Hold returns.

Comparison	DM stat	p -Value	Result
CGX vs. BTC B and H	1.005	0.315	Fail to reject H_0

The DM statistic of 1.005 ($p = 0.315$) indicates that the mean return difference between CGX and Buy and Hold is not statistically significant at the 5% level. This is consistent with the bootstrap analysis and reinforces CGX's positioning as a risk-reduction system: the mean return is not the system's value proposition; the tail-risk reduction and drawdown protection are.

To directly test the downside protection hypothesis, we extend the DM framework using an asymmetric loss function $L(r_t) = \max(-r_t, 0)$ that penalizes negative returns only, ignoring positive returns entirely. This formulation tests whether CGX achieves statistically significant downside loss reduction relative to Buy and Hold (Table 17):

Table 17. Diebold–Mariano tests under symmetric and asymmetric loss functions.

Comparison	DM Stat	<i>p</i> -Value	Result
CGX vs. BTC (mean return)	1.005	0.315	Fail to reject
CGX vs. BTC (downside loss)	10.793	$<10^{-6}$	Reject H_0

The contrast is stark: while mean returns are statistically indistinguishable ($p = 0.315$), the downside loss differential is overwhelmingly significant ($DM = 10.793$, $p < 10^{-6}$). The mean downside loss per biweekly period is 1.49 percentage points lower for CGX than for Buy and Hold. This result precisely quantifies CGX's value proposition: statistically insignificant return differences paired with highly significant risk reduction. The system's contribution is not in predicting market direction but in systematically avoiding the worst outcomes through regime-adaptive exposure management.

6.12. Bootstrap Inference

To assess the statistical significance of performance differences, we construct nonparametric bootstrap confidence intervals from the pooled 413 biweekly return observations using 10,000 resamples with replacement.

We construct block bootstrap confidence intervals (block size $b = 13$, 10,000 resamples) to account for time-series dependence [31].

Mean return (H_0^μ): Block bootstrap 95% CI = $[-0.69\%, +0.18\%]$, containing zero. Non-significance is expected: CGX is designed to reduce risk, not enhance mean returns. A formal Sharpe ratio comparison is in Section 6.13.

Volatility (H_0^σ): CGX biweekly std = 0.72% vs. BTC = 5.12% (86% reduction). Block bootstrap 95% CI = $[-4.94\%, -3.88\%]$, excluding zero—the most statistically robust result in the paper.

CVaR (H_0^{CVaR}): CGX CVaR(5%) = -1.53% vs. BTC = -11.12% (86% reduction). Bootstrap CI excludes zero, confirming statistically significant tail-risk mitigation.

6.13. Sharpe Ratio Test and Sortino Analysis

Sharpe ratio test. The Jobson–Korkie test [26] yields $z = 3.277$ ($p = 0.001$); however, this assumes i.i.d. normal returns—violated for cryptocurrency. The bootstrap 95% CI for the Sharpe difference is $[-0.181, +0.311]$, containing zero. We rely on the bootstrap: Sharpe differences remain statistically uncertain, consistent with our positioning as a risk-reduction rather than return-maximization system.

Sortino ratio. The Sortino ratio penalizes only downside deviation, matching CGX's capital preservation objective. CGX downside volatility is 3.38% annualized versus BTC's 25.09% (87% reduction), yielding Sortino ratios of 0.363 vs. 0.274. The economically meaningful result is the $9\times$ reduction in tail loss severity: CGX's worst 5% of periods average -0.86% versus -7.59% for Buy and Hold.

6.14. VaR Backtesting (Kupiec POF Test)

To formally validate CGX’s tail-risk model, we apply the Kupiec Proportion of Failures (POF) test [32]. Let x denote the observed number of exceedances (returns below the $\alpha = 5\%$ VaR), T the total observations, and $\hat{p} = x/T$ the observed failure rate. The null hypothesis $H_0 : \hat{p} = \alpha$ is tested via:

$$LR_{\text{POF}} = -2 \ln \left[\frac{\alpha^x (1 - \alpha)^{T-x}}{\hat{p}^x (1 - \hat{p})^{T-x}} \right] \sim \chi^2(1) \quad (9)$$

System	VaR(5%)	Exceedances	Expected	p-Value	Result
CGX	−0.86%	21/413 (5.1%)	20.7	0.937	Pass
BTC B&H	−7.59%	21/413 (5.1%)	20.7	0.937	Pass

Both systems pass the Kupiec test ($p = 0.937$): the observed exceedance rate of 5.1% is statistically indistinguishable from the expected 5%, confirming that the empirical VaR model is correctly specified for both. The critical distinction lies in the magnitude of the VaR threshold: CGX’s 5% VaR is −0.86% per biweekly period versus −7.59% for Buy and Hold—an 89% reduction in tail loss severity. The consensus gate does not change the frequency of adverse events but dramatically reduces their size, consistent with the system’s regime-avoidance mechanism rather than per-trade prediction.

6.15. Cross-Asset Applicability: ETH-USD

To assess whether CGX’s architectural principles generalize beyond Bitcoin, we examine the ETH-USD market over the same 2022–2025 period. While we do not run the full CGX system on ETH (which would require 400+ additional LLM debates), we characterize the market conditions that CGX would encounter:

	2022	2023	2024	2025
ETH Return	−68.1%	+95.9%	+41.7%	−10.8%
ETH Ann. Vol	87.0%	46.4%	64.3%	74.7%
ETH MaxDD	74.1%	27.4%	45.3%	60.1%

BTC-ETH daily return correlation: $\rho = 0.84$

ETH exhibits qualitatively similar regime structure to BTC (bear in 2022, bull in 2023–2024, mixed in 2025) but with higher volatility (46–87% annualized vs. BTC’s 20–32%) and deeper drawdowns (74.1% vs. 66.5% in 2022). The high BTC-ETH correlation ($\rho = 0.84$) suggests that a Bear agent trained on ETH would exhibit similar regime-gating behavior, but the higher volatility would amplify both the protection benefit (larger drawdowns to avoid) and the opportunity cost (larger rallies to miss). To move beyond characterization, we conduct a cross-asset transfer experiment that tests the mechanism CGX actually claims to provide: regime-level exposure management rather than asset-specific direction prediction. We take the session-level exposure sequence produced by the BTC debates—the recorded output of the consensus gate at each of the 417 session dates—and apply it, unchanged, to ETH-USD returns over the identical dates, with the same 0.1% fee charged on exposure changes. Under this transfer, the gate’s regime signal is tested on an asset it never analyzed. Table 18 reports the results.

Table 18. ETH transfer test: BTC-derived CGX exposure applied to ETH-USD.

Year	CGX Exposure Transfer			ETH Buy and Hold		
	Ret.	MaxDD	Vol.	Ret.	MaxDD	Vol.
2022	−10.2%	11.3%	3.1%	−68.1%	71.6%	43.3%
2023	+11.4%	5.9%	4.1%	+93.3%	25.9%	22.8%
2024	+10.2%	7.7%	5.0%	+42.7%	43.6%	34.0%
2025	+3.7%	11.5%	5.4%	−15.0%	58.7%	33.3%
2022–2025	+15.2%	12.6%	4.5%	−22.0%	71.6%	34.1%

Returns and drawdowns computed on ETH-USD closes sampled at the 417 BTC session dates; annualized volatility from biweekly returns. Buy-and-hold figures therefore differ slightly from the calendar-year values in the characterization table above.

The transferred gate preserves its protective character on an asset it never analyzed: maximum drawdown falls from 71.6% to 12.6% and annualized volatility from 34.1% to 4.5% over the full period. Strikingly, on ETH the transfer also wins on absolute return (+15.2% vs. −22.0% cumulative), because ETH—unlike BTC—ended the four-year window below its starting price, so the opportunity cost of abstention was small while the protection benefit was large. This asymmetry is exactly what the regime-management interpretation predicts: the more adverse the asset’s regime mix, the more the consensus gate contributes. Two caveats bound this result. First, the exposure sequence was generated from BTC market data; the transfer is informative because of the high BTC–ETH regime correlation ($\rho = 0.84$), and per-asset debates could plausibly perform better (or reveal asset-specific failure modes). Second, ETH remains a large-cap cryptocurrency whose regimes co-move with BTC; transfer to small-cap altcoins, equities, or commodities is untested. ETH-native debate validation (~400 additional LLM sessions) and non-cryptocurrency assets remain important directions for future work.

6.16. Bear Agent Interpretability

A key question is whether the Bear agent’s blocking behavior is directionally correct—i.e., whether blocked sessions would have produced negative BTC returns. We analyze the mean BTC return during blocked versus passed sessions:

Year	N_{block}	N_{pass}	$\text{Ret}_{\text{block}}$	Ret_{pass}	Correct?
2022	86	7	−0.45%	−1.99%	No
2023	13	75	+1.93%	+0.41%	No
2024	11	80	+1.22%	+0.49%	No
2025	31	63	−0.23%	+0.14%	Yes

In three of four years, blocked sessions had higher average BTC returns than passed sessions, meaning the Bear agent was not directionally accurate in its blocking. In 2022, the Bear blocked 86 of 93 sessions, but the 7 passed sessions actually had worse returns (−1.99%) than the blocked ones (−0.45%). Only in 2025 was blocking directionally correct.

This counterintuitive result reinforces the calibration finding (Section 6.10): the Bear agent does not predict direction. Its value lies in wholesale market avoidance during bear regimes. In 2022, the Bear blocked 93% of sessions—it did not matter that individual blocked sessions had slightly better returns than passed ones, because the aggregate effect of near-total non-participation during a −64% market decline produced a −9.45% loss versus −64.19% for Buy and Hold. The mechanism is regime identification (correctly identifying an adverse environment) rather than session-level prediction (correctly timing individual periods within that regime).

6.17. Cumulative Compounding

Starting from USD 100,000 on 1 January 2022 and compounding annual returns:

$$\begin{aligned} V_{\text{CGX}} &= 100,000 \times 0.9055 \times 1.1765 \times 1.1454 \times 0.9851 \\ &= \text{USD } 120,204 \quad (+20.20\%) \end{aligned} \quad (10)$$

$$\begin{aligned} V_{\text{BTC}} &= 100,000 \times 0.3581 \times 2.5529 \times 2.1003 \times 0.8995 \\ &= \text{USD } 172,711 \quad (+72.71\%) \end{aligned} \quad (11)$$

The four-year CAGR is 4.71% for CGX versus 14.64% for BTC. The gap reflects the system's fundamental trade-off: trading 50 percentage points of cumulative return for an 85% reduction in peak volatility, a 7× reduction in maximum drawdown, and a 2.1× better Calmar ratio.

7. Discussion

7.1. Why Adversarial Debate Suits Financial Markets

Financial markets function because participants with opposing views trade with each other; prices aggregate diverse perspectives that no single participant possesses. CGX institutionalizes this dynamic: the Bull Agent surfaces underappreciated opportunities while the Bear Agent identifies overpriced risk. The empirical finding that consensus never exceeds 52% in the aggregation study validates this design—genuinely opposed analytical mandates maintain productive disagreement that collaborative architectures may not achieve.

7.2. The Bear Gate as Implicit Regime Detector

The multi-year validation reveals what appears to be an emergent property not present by design: the Bear agent functions as an implicit market regime classifier. Without any conditioning on labeled regimes, the LLM produces Bear scores of −12.0 during the 2022 crash (93% block rate) and −5.8 to −6.0 during the 2023–2024 bull markets (12–15% block rate). This gradient is consistent with the LLM's ability to synthesize macroeconomic context, on-chain signals, and sentiment data in a way that traditional quantitative regime indicators—which typically require parameter tuning—may not easily replicate.

The practical implication is that regime-aware position sizing emerges for free from the adversarial debate structure: no explicit regime model is required when the Bear agent's adversarial mandate already penalizes bullish arguments proportional to systemic risk.

This emergent regime sensitivity can be linked to the GARCH volatility dynamics established in Section 4. The GARCH(1,1) model identifies a persistence parameter of $\alpha_1 + \beta_1 = 0.979$, implying that volatility shocks persist for approximately 32 trading days. The Bear agent's scores exhibit a qualitatively similar persistence: the mean Bear score transitions gradually from −12.0 in 2022 to −5.8 in 2023 (a shift that tracks the declining conditional variance as the market recovered from the FTX-era volatility spike). The LLM appears to encode the same volatility-clustering dynamics that the GARCH model captures parametrically, but through contextual analysis of news, sentiment, and market narratives rather than through autoregressive variance modeling. Whether this is a genuine causal mechanism or a coincidental correlation remains an open empirical question that would require controlled experiments with synthetic market data.

7.3. Aggregation Matters as Much as Debate

Our most significant finding from the aggregation study is that individual agents produce rich, informative signals (87% bull-buy, 59% bear-sell, meaningful confidence variation), but the Meta-Evaluator's conservative synthesis collapses this into universal

holds. The aggregation mechanism is thus as consequential as the debate protocol itself. This has implications for multi-agent system design broadly: the standard approach of appointing a single arbiter may systematically discard valuable information embedded in agent-level disagreement patterns.

7.4. Positioning Relative to Published LLM-Trading Results

Direct numerical comparison with published agentic trading systems is complicated by a structural problem in this literature: the studies rarely share an asset universe or evaluation period (FinMem and TradingGroup, which use the same five equities and window, are the exception). Published evaluations are predominantly single-regime and equity-focused, and cover windows of two to seven months. Table 19 juxtaposes the headline results reported by each framework’s authors with CGX’s, and should be read as context, not ranking: the numbers are not commensurable, and in at least one case the original authors themselves flag an anomaly (TradingAgents reports a Sharpe ratio of 8.21 on a three-month window and notes that it “exceeds our expected empirical range”).

Three structural observations survive the incomparability caveat. First, FinAgent’s ETH-USD result is the only cryptocurrency evaluation among the six frameworks, and it covers a seven-month window; CGX’s evaluation is, to our knowledge, the only multi-year, multi-regime cryptocurrency evaluation of an LLM trading agent, including a −64% bear market—the regime in which risk mechanisms are actually tested. Second, judged by the maximum drawdowns the six papers themselves report (0.9–12.7%), none of these evaluations encountered—let alone survived—a deep drawdown regime, whereas CGX’s window contains BTC’s 66.5% peak-to-trough collapse of 2022. Third, the comparison axis on which CGX is designed to compete—worst-case drawdown across regimes—is one that two-to-seven-month evaluation windows cannot measure. We therefore view Table 19 together with the mechanism comparison of Table 1 as jointly answering the positioning question: CGX occupies a design point (consensus-gated abstention, multi-regime validation, capital preservation) that the current literature reports neither results for nor mechanisms against.

Table 19. Published headline results of LLM trading frameworks (as reported by their authors) vs. CGX.

Framework	Asset (Headline)	Period	Backbone	Return	Sharpe	MaxDD
FinMem [3]	TSLA (equity)	October 2022–April 2023	GPT-4-Turbo	61.8%	2.68	10.8%
FinAgent [4]	ETH-USD	June 2023–January 2024	GPT-4(V)	43.1%	1.18	12.7%
TradingAgents [1]	AAPL (equity)	January–March 2024	o1-preview	26.6%	8.21 [†]	0.9%
TradingGroup [2]	AMZN (equity)	October 2022–April 2023	GPT-4o-mini	40.5%	3.86	2.1%
CGX (aggregation study)	BTC/USDT	January–December 2024	Claude Sonnet 4.6	58.0%	1.90	11.6%
CGX (multi-regime)	BTC/USDT	2022–2025 (4 years)	Claude Sonnet 4.6	20.2%	0.34	10.2%

[†] The TradingAgents authors state this value “exceeds our expected empirical range”.

7.5. Capital Preservation vs. Return Maximization

As established in the introduction, CGX is a capital preservation architecture, not an alpha-maximizing strategy. The four-year results confirm this positioning: the Bear gate consistently mitigates catastrophic drawdowns (85% reduction in 2022), at the cost of 52 percentage points of cumulative return foregone relative to Buy and Hold—the expected price of spending 87.4% of the time in cash.

Viewed through a portfolio lens, CGX’s return stream—low beta, bounded drawdowns, regime-adaptive exposure—resembles a tail-risk hedge more than a directional strategy. Its value is best measured by the risk budget it frees for allocation to higher-returning strategies within a diversified portfolio, rather than by standalone absolute return.

7.6. The Threshold Filtering Effect

The non-monotonic relationship between threshold and return (Majority $\theta = 0.35$ outperforming $\theta = 0.00$) reveals a quality-filtering dynamic analogous to optimal stopping problems in decision theory. Moderate thresholds remove low-conviction signals that would generate losing trades, concentrating capital on high-quality opportunities identified through genuine agent agreement.

7.7. Robustness Considerations

Several factors warrant caution when interpreting the reported results. First, the threshold $\theta = 0.35$ is a hyperparameter selected on the basis of 2024 performance. Its optimality is specific to the observed consensus score distribution, which clusters in the $[0.22, 0.52]$ range across all 52 debates. The non-monotonic return profile across thresholds—where $\theta = 0.35$ outperforms both $\theta = 0.00$ and higher values—can be understood through the lens of signal detection theory. At $\theta = 0.00$, all agent-agreement signals are acted upon regardless of conviction quality, admitting noise. At $\theta = 0.35$, only signals in the upper portion of the empirical consensus distribution pass the filter, analogous to setting a detection threshold that balances Type I errors (false signals acted upon) against Type II errors (true signals missed). The walk-forward analysis (Section 6) provides partial out-of-sample validation, but we acknowledge that a different consensus distribution—arising, for example, from a different LLM or prompt structure—could shift the optimal threshold. The consensus function parameters α and β introduce additional sensitivity: at $\alpha = \beta = 1$ (the current specification), agreement and evidence quality contribute equally. Increasing α relative to β would make the system more sensitive to agent disagreement, potentially increasing the block rate and further reducing both returns and drawdowns. A systematic grid search over (θ, α, β) is computationally prohibitive given the cost of LLM-generated debates ($\sim$ USD 0.40 per session), which is itself a structural limitation of the experimental design.

To assess the sensitivity of results to threshold selection, Table 20 reports CGX performance across a range of θ values applied to the full four-year dataset using Majority Vote aggregation. The results demonstrate that performance degrades gracefully rather than exhibiting cliff-like sensitivity to threshold choice:

Table 20. Threshold sensitivity (Majority Vote, 2022–2025).

θ	Return	MaxDD	Sharpe	Calmar	Trades
0.00	+15.8%	12.4%	0.31	0.30	196
0.20	+18.1%	11.2%	0.38	0.38	142
0.30	+19.4%	10.5%	0.42	0.43	108
0.35	+20.2%	10.2%	0.43	0.46	96
0.40	+17.6%	9.8%	0.40	0.42	71
0.50	+8.3%	6.1%	0.30	0.32	28

Several observations reduce overfitting concerns. First, performance varies smoothly across the $\theta \in [0.20, 0.40]$ range, with Calmar ratios between 0.38 and 0.46—all substantially above Buy and Hold's 0.22. Second, maximum drawdown monotonically decreases with higher θ (from 12.4% at $\theta = 0.00$ to 6.1% at $\theta = 0.50$), confirming that the consensus gate consistently provides drawdown protection regardless of the exact threshold. Third, the optimal $\theta = 0.35$ is not an isolated peak but falls within a broad plateau of near-optimal performance.

The risk of post hoc selection bias in θ merits explicit acknowledgment. In the aggregation study (Section 5), the full set of $\{\theta, \text{strategy}\}$ combinations was evaluated on the same 2024 data, and the best-performing configuration was reported. This constitutes a multiple-

comparisons problem: with six strategies and multiple thresholds, some configuration will appear optimal by chance. The walk-forward analysis (2022–2023 in-sample, 2024–2025 out-of-sample) partially mitigates this concern, and the observation that out-of-sample performance exceeds in-sample performance provides evidence against pure overfitting. Nevertheless, for a production deployment, we recommend rolling threshold optimization with a recalibration window of 6–12 months, re-estimating θ from the empirical consensus distribution on a trailing basis. An adaptive θ that tracks the moving percentile of the consensus score distribution (e.g., the 70th percentile) would further reduce sensitivity to distributional shift.

Second, the system’s performance is inherently regime-dependent: CGX generates positive alpha in bear and mixed markets (2022, 2025) while substantially underperforming in strong bull markets (2023, 2024). In a hypothetical future dominated by sustained upward trends, the Bear gate’s protective mechanism would impose persistent opportunity cost. The four-year evaluation captures two bull and two bear/mixed years, providing a balanced but not exhaustive sample of regime diversity.

Third, all backtest results are subject to the standard limitations of historical simulation, including the absence of slippage, market impact, and counterparty risk. While these factors are less consequential for the biweekly trading frequency employed by CGX than for higher-frequency strategies, they would need to be quantified in any live deployment.

Fourth, the biweekly return series exhibits time-series dependence, violating the i.i.d. assumption underlying standard inference. While the ARIMA and GARCH analyses (Section 4) characterize the autocorrelation and heteroscedasticity structures, the bootstrap confidence intervals reported above use block resampling implicitly through the temporal pooling of returns; a stationary block bootstrap [31] would provide more conservative intervals. With 413 biweekly observations across four years, the effective sample size for annual-frequency conclusions is only $N_{\text{eff}} = 4$, severely limiting the power of year-level statistical tests. We therefore interpret multi-year results as descriptive evidence of regime-adaptive behavior rather than statistically conclusive proof of superiority.

7.8. Input Integrity and Adversarial Robustness

CGX’s decision quality is bounded by the integrity of its inputs. Deployed in an open network environment, the agents consume technical indicators computed from exchange feeds together with textual market context and news summaries; future multimodal extensions would additionally ingest charts, social-media imagery, and video. Each channel is a potential attack surface. Three threat classes are salient. (i) Indirect prompt injection: Adversarial instructions embedded in retrieved news or web content can hijack an LLM agent’s behavior without access to the model itself, a vulnerability demonstrated across production LLM-integrated applications [33]. (ii) False-narrative injection: Coordinated misinformation campaigns—increasingly cheap to mount with AI-generated text, images, and deepfake video of public figures—can manufacture the very “market context” that agents are asked to reason over, biasing Bull or Bear evidence without touching the price feed. Detecting such manipulated media is an active research field [34], and integrating multimodal forgery detectors as input filters is a natural extension for any multimodal trading agent. (iii) Data-feed tampering: Man-in-the-middle modification of market data or multimedia streams in transit; here, authenticated and encrypted transport is the baseline defense, and the multimedia-security literature provides both attacks and remedies for protecting visual data streams [35,36].

Several mitigations map directly onto the CGX architecture. Input provenance should be enforced at ingestion (signed feeds, allow-listed sources), and evidence derived from unverifiable or single-source content can be capped in the Meta-Evaluator’s quality score

E_q , which by Equation (3) proportionally suppresses consensus and thus position size. Cross-source corroboration—requiring independent confirmation before an evidence item receives full weight—raises the cost of narrative injection. Notably, the consensus gate itself provides a structural, if partial, defense: a fabricated bullish narrative must survive cross-examination by a Bear agent whose mandate is to attack exactly such claims, and manipulation that convinces only one side increases inter-agent disagreement, driving C toward zero and blocking execution. The gate therefore fails safe under one-sided input manipulation; a coordinated attack would need to deceive structurally opposed agents simultaneously. Systematic red-teaming of the debate protocol under adversarial inputs—and quantifying how much injected evidence is required to flip a gate decision—is an important direction for future work.

7.9. Limitations

Single debate asset. All LLM debates were run on BTC/USDT market data. The ETH evidence in Section 6.15 is an exposure-transfer analysis—it shows that the BTC-derived gate signal protects an ETH portfolio, not that asset-native ETH debates would behave identically. Validation with asset-native debates across altcoins, equities, and cross-asset portfolios is necessary to establish whether the consensus-gating mechanism generalizes beyond large-cap cryptocurrency regimes.

Long-only constraint. The Bear Agent’s signals are used only to reduce or close positions, not to initiate short positions. This architectural choice means that CGX cannot profit from correctly identified downturns; a short-selling extension could meaningfully improve bear-market performance beyond the -9.45% recorded in 2022.

Post hoc threshold selection. The optimal $\theta = 0.35$ was identified retrospectively on 2024 data. While the walk-forward analysis provides partial mitigation, formal walk-forward optimization with rolling recalibration is required for deployment to avoid residual look-ahead bias.

Correlated agent outputs. All agents share the same underlying LLM (Claude Sonnet 4.6), violating the conditional independence assumption (A1) of Proposition 1. This correlation manifests at multiple levels: (i) Knowledge correlation—All agents share identical training data, producing correlated blind spots for events outside the training distribution; (ii) Reasoning correlation—The same model architecture and weights produce structurally similar inference patterns, even under opposing prompts; (iii) Calibration correlation—Confidence scores are generated by the same calibration function, producing systematically similar uncertainty estimates. The adversarial role assignment mitigates but does not eliminate these correlations: empirically, Bull and Bear agents agree on direction in 41% of debates (16/39), suggesting a residual correlation of approximately $\rho \approx 0.2\text{--}0.4$, which reduces the effective agent count from $n = 3$ to $n_{\text{eff}} \approx 1.5\text{--}2.1$ (cf. Proposition 1). A heterogeneous agent architecture—deploying distinct LLMs (e.g., GPT-4 for Bull, Gemini for Bear, an open-source model for Risk) with different training corpora and reasoning patterns—would reduce inter-agent correlation and increase the diversity benefit. The trade-off is system complexity: heterogeneous agents may produce inconsistent output formats, varying quality levels, and incomparable confidence scales, requiring additional calibration infrastructure. We view this as a promising direction for future work, particularly as the LLM ecosystem diversifies.

Engineering pathways to heterogeneous ensembles. Replacing individual agents with heterogeneous open-source models—for example, Llama-family models [37] as the Bear agent or DeepSeek-R1 [38] as the Risk agent—raises four concrete engineering challenges beyond the architectural motivation discussed above. (i) Output-format alignment: CGX requires every agent to emit the structured tuple of Equation (2) as schema-validated JSON.

Frontier APIs enforce this natively, whereas open-source models exhibit higher format-violation rates; in self-hosted deployments this is addressable with grammar-constrained or schema-guided decoding, which restricts generation to outputs that parse against the tuple schema by construction. (ii) Confidence-scale alignment: The consensus function (Equation (3)) compares signed convictions $d_i = \sigma(r_i) \cdot c_i$ across agents, implicitly assuming that a confidence of, say, 0.7 means the same thing for every agent. Different model families are differently calibrated, so raw confidences from heterogeneous backbones are not directly commensurable. A practical remedy is per-model post hoc calibration—fitting an isotonic or Platt mapping on a held-out set of debates—or, more simply, rank-normalizing each agent’s confidence through its own empirical distribution before computing $\text{Var}(\{d_i\})$, so that the variance term measures genuine disagreement rather than inter-model calibration offsets. (iii) Evidence-quality comparability: E_q is currently scored by a single Meta-Evaluator; with heterogeneous debaters the Meta-Evaluator should remain a single fixed judge so that evidence scores stay on one scale, at the cost of re-introducing a shared-model dependency at the synthesis stage. (iv) Capability asymmetry: Adversarial debate is only as strong as its weakest critic—a weaker open-source Bear agent would systematically lose cross-examinations against a stronger Bull agent, biasing consensus upward precisely when skepticism matters most. Role assignment should therefore match model capability to role criticality, and capability-weighted convictions offer a further safeguard. These are engineering rather than conceptual obstacles, and the potential payoff—decorrelated errors across training corpora and architectures—directly targets the effective-agent-count reduction ($n_{\text{eff}} \approx 1.5\text{--}2.1$) quantified above.

Stochastic non-reproducibility. LLM outputs are inherently stochastic: the same market data and prompt may produce different agent forecasts across runs due to sampling temperature and internal randomness. Our experiments use a single realization of each debate; repeated trials would yield a distribution of outcomes rather than a point estimate. The reported results should therefore be understood as one draw from this distribution. We estimate, based on limited replication experiments, that session-level stochasticity introduces approximately $\pm 2\text{--}3$ percentage points of uncertainty in annual return estimates. Quantifying this variance rigorously through repeated full-system replications is computationally expensive ($\sim$ USD 170 per four-year run) but represents an important direction for future work.

Lack of directional predictability. As established in the calibration analysis (Section 6.10), the consensus score carries no per-trade directional information ($\rho = 0.005$, $p = 0.94$) and no calibration technique recovers a useful probability mapping. This is consistent with the Adaptive Markets Hypothesis [19]. CGX’s contribution is regime-level exposure management, not directional prediction.

Meta-Evaluator failure. The Meta-Evaluator produced zero trades across all 52 debates at every threshold tested, including $\theta = 0.20$. While we interpret this as evidence of structural collapse prevention, it also represents a design limitation: the designated synthesis component fails entirely at its primary function. The Meta-Evaluator’s conservative bias may reflect the LLM’s tendency toward hedged, non-committal synthesis when presented with genuinely conflicting evidence—a known property of instruction-tuned models [28]. This failure motivated the alternative aggregation strategies that ultimately drove the positive results.

Backtest fidelity. All results assume instantaneous execution at observed prices with fixed 0.1% transaction costs and no market impact. While BTC/USDT’s deep liquidity ($>$ USD 1B daily volume on major exchanges) makes market impact negligible for the position sizes employed ($\sim$ USD 12,600 average), latency between debate completion

(~15 min) and execution could introduce adverse price movement. These effects are not modeled.

Cost and latency. At ~ USD 0.40 per debate and ~15 min per cycle, CGX is suited to position-level trading on daily or longer timeframes rather than intraday strategies.

7.10. Ethical Considerations

Autonomous AI agents managing capital raise questions regarding accountability for losses, market manipulation risk from widespread deployment, systemic risk from correlated AI-driven behavior, and concentration of financial decision-making in LLM providers. CGX addresses accountability through comprehensive logging of all debate transcripts and decision rationale.

8. Conclusions

We present CGX, a multi-agent LLM architecture for cryptocurrency trading validated through two complementary experiments: a 52-week aggregation study and a four-year multi-regime evaluation spanning 417 sessions and 196 trades. Statistical evidence supports its capital-preservation positioning: mean return differences are not statistically significant (DM $p = 0.315$), while risk reduction metrics—86% volatility reduction, 86% CVaR reduction, 85% drawdown reduction—are statistically robust (block bootstrap CIs exclude zero; downside DM $p < 10^{-6}$).

Four principal findings emerge. First, adversarial role assignment structurally prevents debate collapse, with consensus scores bounded between 22 and 52% across all tested periods. Second, consensus thresholds function as a tunable risk–return dial: the best configuration achieves Sharpe 1.90 with 11.6% drawdown, a $3\times$ improvement over trend following and 80% drawdown reduction versus Buy and Hold. Third, the signal aggregation strategy matters as much as the debate protocol itself—the Meta-Evaluator’s conservative synthesis discards profitable information that alternative aggregation recovers. Fourth, the Bear consensus gate exhibits behavior consistent with emergent regime detection: blocking 93% of sessions during the 2022 bear market and only 12% during the 2024 bull run, without any explicit regime conditioning—a property that appears to emerge naturally from the LLM’s adversarial analytical mandate.

Over four years (2022–2025), CGX compounds to USD 120,204 (+20.2%) versus USD 172,711 (+72.7%) for Buy and Hold, with a Calmar ratio of 0.462 versus 0.220—a $2.1\times$ improvement. Maximum drawdown never exceeds 10.2%, while BTC experienced a 66.5% drawdown in 2022 alone. The calibration analysis (Section 6.10) confirms that CGX’s edge lies in when not to trade rather than directional prediction—a mechanism that connects to classical abstention principles in decision theory and may generalize beyond financial trading to any domain where the cost of incorrect action exceeds the cost of inaction.

Future work includes multi-asset validation with asset-native debates (extending the ETH exposure-transfer result to ETH-native debates, and to equity indices), walk-forward threshold optimization, short-selling extension for bear markets, heterogeneous agent architectures using distinct LLMs to reduce inter-agent correlation, learned aggregation functions via meta-learning, and live forward testing with constrained capital. Quantifying the variance of system performance across independent debate replications is particularly important for establishing the robustness of the reported risk-reduction properties.

Author Contributions: Conceptualization, A.B. and F.M.; methodology, A.B., A.K. and F.M.; software, A.B.; validation, A.B., A.K. and F.M.; formal analysis, A.B. and F.M.; investigation, A.B.; resources, A.B., A.K. and F.M.; data curation, A.B.; writing—original draft preparation, A.B.; writing—review and editing, A.B., A.K. and F.M.; visualization, F.M.; supervision, F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The R source code, experimental scripts, datasets, and supporting materials used in this study are available from the corresponding author upon reasonable request. These materials include the CGX v2.7 implementation, the datasets used in the experiments, the statistical analysis scripts, and the deep-learning baseline and cross-asset transfer experiments with fixed random seeds and archived results. Market data were obtained from publicly available sources (Binance BTC/USDT; Yahoo Finance BTC-USD and ETH-USD).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xiao, Y.; Sun, E.; Luo, D.; Wang, W. TradingAgents: Multi-Agents LLM Financial Trading Framework. *arXiv* **2024**, arXiv:2412.20138. Available online: <https://arxiv.org/abs/2412.20138> (accessed on 25 March 2026).
2. Tian, F.; Salim, F.D.; Xue, H. TradingGroup: A Multi-Agent Trading System with Self-Reflection and Data-Synthesis. *arXiv* **2025**, arXiv:2508.17565. Available online: <https://arxiv.org/abs/2508.17565> (accessed on 25 March 2026).
3. Yu, Y.; Li, H.; Chen, Z.; Jiang, Y.; Li, Y.; Suchow, J.W.; Zhang, D.; Khashanah, K. FinMem: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design. *IEEE Trans. Big Data* **2025**, *11*, 3443–3459. [\[CrossRef\]](#)
4. Zhang, W.; Zhao, L.; Xia, H.; Sun, S.; Sun, J.; Qin, M.; Li, X.; Zhao, Y.; Zhao, Y.; Cai, X.; et al. A Multimodal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024; pp. 4314–4325. [\[CrossRef\]](#)
5. Li, W.W.; Kim, H.; Cucuringu, M.; Ma, T. Can LLM-Based Financial Investing Strategies Outperform the Market in Long Run? In Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26), Jeju, Republic of Korea, 9–13 August 2026; pp. 2711–2722. [\[CrossRef\]](#)
6. Kaesberg, L.B.; Becker, J.; Wahle, J.P.; Ruas, T.; Gipp, B. Voting or Consensus? Decision-Making in Multi-Agent Debate. In *Findings of the Association for Computational Linguistics: ACL 2025*; ACL: Stroudsburg, PA, USA, 2025; pp. 11640–11671. [\[CrossRef\]](#)
7. Cui, Y.; Fu, H.; Zhang, H.; Wang, L.; Zuo, C. Free-MAD: Consensus-Free Multi-Agent Debate. In *Findings of the Association for Computational Linguistics: ACL 2026*; ACL: Stroudsburg, PA, USA, 2026; pp. 31977–31997. [\[CrossRef\]](#)
8. Hu, T.; Tan, Z.; Wang, S.; Qu, H.; Chen, T. Multi-Agent Debate for LLM Judges with Adaptive Stability Detection. *NeurIPS* **2026**, *38*, 46504–46540.
9. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive Mixtures of Local Experts. *Neural Comput.* **1991**, *3*, 79–87. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Krogh, A.; Vedelsby, J. Neural Network Ensembles, Cross Validation, and Active Learning. *Adv. Neural Inf. Process. Syst.* **1995**, *7*, 231–238.
11. Kuncheva, L.I.; Whitaker, C.J. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Mach. Learn.* **2003**, *51*, 181–207. [\[CrossRef\]](#)
12. Chow, C.K. On Optimum Recognition Error and Reject Tradeoff. *IEEE Trans. Inf. Theory* **1970**, *16*, 41–46. [\[CrossRef\]](#)
13. El-Yaniv, R.; Wiener, Y. On the Foundations of Noise-Free Selective Classification. *J. Mach. Learn. Res.* **2010**, *11*, 1605–1641.
14. Irving, G.; Christiano, P.; Amodei, D. AI Safety via Debate. *arXiv* **2018**, arXiv:1805.00899. Available online: <https://arxiv.org/abs/1805.00899> (accessed on 25 March 2026).
15. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J.B.; Mordatch, I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, 21–27 July 2024.
16. Kelly, J.L. A New Interpretation of Information Rate. *Bell Syst. Tech. J.* **1956**, *35*, 917–926. [\[CrossRef\]](#)
17. Sunstein, C.R.; Hastie, R. *Wiser: Getting Beyond Groupthink to Make Groups Smarter*; Harvard Business Review Press: Boston, MA, USA, 2015.
18. Nickerson, R.S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psycho.* **1998**, *2*, 175–220. [\[CrossRef\]](#)
19. Lo, A.W. The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective. *J. Portf. Manag. Forthcom.* **2004**, *30*, 15–29. [\[CrossRef\]](#)
20. Bollerslev, T. Generalized Autoregressive Conditional Heteroscedasticity. *J. Econom.* **1986**, *31*, 307–327. [\[CrossRef\]](#)
21. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2016), San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
22. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Cho, K.; Van Merriënboer, B.; Gulçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734.

24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *NeurIPS* **2017**, *30*, 5998–6008.
25. Lo, A.W. The Statistics of Sharpe Ratios. *Financ. Anal. J.* **2002**, *58*, 36–52. [[CrossRef](#)]
26. Jobson, J.D.; Korkie, B.M. Performance Hypothesis Testing with the Sharpe and Treynor Measures. *J. Finance* **1981**, *36*, 889–908. [[CrossRef](#)]
27. Ledoit, O.; Wolf, M. Robust Performance Hypothesis Testing with the Sharpe Ratio. *J. Empir. Finance* **2008**, *15*, 850–859. [[CrossRef](#)]
28. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [[CrossRef](#)]
29. Barber, B.M.; Odean, T. Trading is hazardous to your wealth: The common stock investment performance of individual investors. *J. Finance* **2000**, *55*, 773–806. [[CrossRef](#)]
30. Diebold, F.X.; Mariano, R.S. Comparing Predictive Accuracy. *J. Bus. Econ. Stat.* **2002**, *20*, 134–144. [[CrossRef](#)]
31. Politis, D.N.; Romano, J.P. The Stationary Bootstrap. *J. Amer. Stat. Assoc.* **1994**, *89*, 1303–1313. [[CrossRef](#)]
32. Kupiec, P.H. Techniques for Verifying the Accuracy of Risk Measurement Models. *J. Deriv.* **1995**, *3*, 73–84. [[CrossRef](#)]
33. Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. Not What You’ve Signed Up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec), Copenhagen, Denmark, 30 November 2023; pp. 79–90.
34. Masood, M.; Nawaz, M.; Malik, K.M.; Javed, A.; Irtaza, A.; Malik, H. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Appl. Intell.* **2023**, *53*, 3974–4026. [[CrossRef](#)]
35. Lin, Y.; Wei, Y.; Chen, D.; Li, Y.; Erkan, U.; Toktas, A.; Gao, S.; Zhang, Y. Cryptanalysis and Improvement of a Video Cryptosystem via Chaos and S-Box. *ACM Trans. Multimed. Comput. Commun. Appl.* **2026**, *22*, 1–286. [[CrossRef](#)]
36. Lin, Y.; Liao, Y.; Zeng, W.; Wei, Y.; Chen, D.; Yuan, X.; Li, Y.; Erkan, U.; Toktas, A.; Zhang, C.; et al. 3D Non-degenerate Hyperchaos: Design, Analysis, and Application in Image Encryption. *IEEE Trans. Consumer Electron.* **2026**, Early Access. [[CrossRef](#)]
37. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783. Available online: <https://arxiv.org/abs/2407.21783> (accessed on 25 March 2026).
38. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning. *Nature* **2025**, *645*, 633–638. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.