Hold-to-Expiry Edge in CFTC-Regulated Crypto Binary Options:

A Calibrated Pricing Model for Kalshi Hourly Range Contracts

OddsReference Research

March 2026

Abstract

We develop and test three pricing models for Kalshi's CFTC-regulated crypto binary options using the Becker (2026) dataset of 72.1 million trades across 877,606 settled contracts. A TWAP-adjusted Black-Scholes model, calibrated via isotonic regression on 5.48 million observations, initially produced +6.7 cents average edge and +221% simulated return across 28,496 qualifying signals. Paper trading revealed a -29.5% return over 37 days, leading to the discovery of three implementation bugs: a subtraction order error affecting 40% of contracts, a time-to-expiry units mismatch causing 7.7x volatility overestimation, and a bracket center-versus-floor interpretation error. After correction, average edge dropped to +1.2 cents per signal (95% CI [+1.15, +1.25], t = 106.47). A probability cap variant showed identical performance (+1.2 cents), while a Kou jump-diffusion model calibrated on 5,152 hourly returns (kurtosis 32.9) improved edge marginally to +1.4 cents. Execution analysis reveals binding constraints: 68% of contracts show zero exit liquidity, volume-edge correlation is 0.849, and the capacity ceiling for a single model-based trader is $5,000-$25,000. The round-trip execution cost of 3.08 cents exceeds the 2.8-3.0 cent gross edge. We conclude that Kalshi crypto binary options are approximately efficiently priced, consistent with our concurrent weather market findings. The paper's methodological contribution is the demonstration that isotonic calibration can simultaneously mask multiple implementation errors, producing plausible backtests from fundamentally broken models.

1. Introduction

Crypto prediction markets have grown rapidly since 2024, with Kalshi processing over $60 million in daily crypto volume by March 2026. Five-minute binary contracts on Bitcoin price ranges account for 73% of this activity, creating the largest regulated crypto derivatives market in the United States. Despite this growth, no published research has systematically tested pricing model efficiency on CFTC-regulated crypto binary options. The intersection of high-frequency crypto price dynamics and regulated binary contract structures creates a unique environment for testing the efficiency of information aggregation in prediction markets.

This paper fills that gap. Using the Becker (2026) dataset -- the most comprehensive publicly available record of Kalshi crypto trading, comprising 72.1 million trades across 877,606 settled range contracts over 385 trading days -- we develop, test, and ultimately fail to profitably trade three pricing models. The failure is itself informative: the market is approximately efficient, and the path to discovering this involved uncovering three implementation bugs that produced a false positive backtest of +221% return. The journey from apparent alpha to confirmed efficiency provides lessons that extend well beyond crypto markets, touching on fundamental questions about model validation, statistical calibration, and the interpretation of backtested returns.

Our research questions are threefold. First, can a TWAP-adjusted pricing model identify tradeable mispricings in Kalshi crypto binary options? Second, how do implementation errors in quantitative models interact with statistical calibration techniques, and what are the diagnostic implications? Third, what are the binding constraints on model-based trading in this market, and do they differ from those observed in traditional options markets?

The paper makes three contributions. First, we provide the first comprehensive efficiency analysis of CFTC-regulated crypto prediction markets, testing three model specifications across nearly 30,000 signals. The scale of the analysis -- 5.48 million calibration observations across 385 trading days -- provides statistical power that is rare in the prediction market literature, where most studies are limited to hundreds or low thousands of contracts. Second, we document a specific and dangerous failure mode: isotonic calibration masking multiple simultaneous implementation bugs, producing +221% backtested returns from a model with three fundamental errors. This finding has implications for any researcher using nonparametric calibration as a validation step, which is standard practice in probabilistic forecasting. Third, we characterize the execution environment -- liquidity distribution, capacity constraints, and slippage budgets -- that determines whether a statistical edge translates to tradeable profit. This microstructure analysis complements Becker's (2026) aggregate statistics with model-specific execution data.

Our findings converge with our concurrent analysis of Kalshi weather prediction markets (OddsReference Research, 2026c), where an NWS ensemble model achieved genuine predictive accuracy but no profitable trading strategies. Different domains, different models, same conclusion: CFTC-regulated prediction markets are approximately efficient when public information is available and events resolve quickly. The convergence across two independent analyses, using fundamentally different data sources and modeling approaches, strengthens the generalizability of the efficiency finding.

The broader context for this research is the rapid maturation of event-driven derivatives in regulated markets. Traditional options markets have been extensively studied for decades, with a well-established literature on pricing efficiency, market microstructure, and the limits of arbitrage (Black and Scholes, 1973; Merton, 1976). Prediction markets -- which price discrete events as binary contracts -- represent a distinct asset class that has received less quantitative attention, particularly in the crypto domain where underlying asset volatility is an order of magnitude higher than equities. Our analysis bridges these literatures by applying established options pricing methodology to a novel market structure, documenting both where the analogy holds and where it breaks down.

The paper proceeds as follows. Section 2 describes the institutional background and market microstructure of Kalshi's crypto product suite. Section 3 presents the data, including the Becker (2026) dataset and our signal extraction methodology. Section 4 develops the three pricing models: TWAP-adjusted Black-Scholes, probability-capped Black-Scholes, and Kou jump-diffusion. Section 5 reports initial (pre-bug-fix) results. Section 6 details the bug discovery process and its methodological implications for calibration-based validation. Section 7 presents corrected results across all three models. Section 8 analyzes execution constraints including liquidity, capacity, and slippage. Section 9 discusses implications for market participants, platform design, and regulation. Section 10 concludes.

2. Institutional Background

2.1 Kalshi Platform

Kalshi operates as a CFTC-designated contract market (DCM), offering binary contracts across event categories including crypto, weather, economics, and politics. The crypto product suite launched in late 2024 with hourly and daily range contracts on BTC and ETH, expanding to include five-minute contracts and additional assets (DOGE, XRP, SOL) by mid-2025. The platform's regulatory status as a DCM distinguishes it from offshore prediction markets such as Polymarket, which operates outside U.S. regulatory jurisdiction. This distinction is material for efficiency analysis: DCM status imposes transparency requirements, position limits, and fee disclosures that reduce information asymmetry between market participants.

Each event produces a strike ladder of approximately 188 contracts. For BTC hourly contracts, strikes are spaced $100 apart (e.g., "BTC above $95,000," "BTC above $95,100," etc.). ETH uses $40 spacing. Each contract is a binary paying $1.00 if the condition is met at settlement, $0.00 otherwise. The strike ladder is centered on the current spot price and

extends approximately 94 strikes in each direction, covering a price range of roughly $9,400 for BTC and $3,760 for ETH. This range captures approximately 99.5% of realized hourly price moves, ensuring that nearly all settlement outcomes fall within the tradeable strike range.

2.2 Settlement Mechanism

Kalshi crypto contracts settle using CF Benchmarks' TWAP (time-weighted average price) methodology. The settlement process samples 60 one-second price readings during the final minute of the contract window, applies a 20% trimmed mean by discarding the 12 highest and 12 lowest readings, and averages the remaining 36. This three-step procedure -- sampling, trimming, averaging -- serves dual purposes: manipulation resistance and variance reduction. Both properties have direct implications for pricing model design.

The variance reduction factor from TWAP settlement can be derived analytically. For N equally-spaced samples from a continuous price process, the variance of the sample mean relative to the instantaneous spot price is given by the ratio:

$$R(N) = (N + 1)(2N + 1) / (6N2)$$

For N = 60 samples: R(60) = (61)(121) / (6 x 3600) = 7381 / 21600 = 0.342. Combined with the 20% trimming, which removes outlier samples and further reduces variance, the effective variance reduction factor is approximately 0.554. This means a pricing model using raw implied volatility will systematically overestimate the probability of large moves near expiry by a factor of 1/0.554 = 1.81. Any model that fails to account for this factor will systematically overprice out-of-the-money contracts and underprice at-the-money contracts in the final minutes before settlement.

Our model implements this as a linear transition: full TWAP adjustment (sigma_effective = sigma x sqrt0.554) for contracts under 1 minute to expiry, no adjustment for contracts beyond 5 minutes, with linear interpolation between. The linear transition reflects the increasing relevance of TWAP mechanics as settlement approaches; beyond 5 minutes, the settlement window is a small fraction of remaining contract life and the adjustment becomes negligible.

The TWAP mechanism provides meaningful manipulation resistance. Moving the settlement price by $100 would require sustaining artificial prices across 36 of 60 sample seconds -- a capital-intensive proposition against the market's resting depth. Kalshi's order book data shows average resting depth of $15,000-$25,000 within 2 cents of the best price, meaning a manipulator would need to absorb approximately $500,000-$900,000 in opposing orders for 36 continuous seconds. This property contributes to market efficiency by discouraging price manipulation near expiry.

2.3 Fee Structure

Taker fees follow a quadratic formula: ceil(0.07 x P x (1 - P) x 100) cents per contract, where P is the transaction price expressed as a probability. Maker fees are 25% of taker fees. The P x (1 - P) structure produces a parabolic fee curve that peaks at 1.75 cents for 50/50 contracts and decreases toward zero at extreme prices. This design has two notable properties: it is revenue-maximizing at the point of maximum uncertainty, and it subsidizes trading at extreme prices where informational content is highest.

The fee structure has distributional consequences for different market participants. A contract priced at 5 cents incurs a taker fee of 0.34 cents (6.8% of contract price), while a contract at 50 cents pays 1.75 cents (3.5%). This asymmetry disproportionately affects traders in the out-of-the-money tails, who are predominantly retail directional bettors. Professional market makers, who transact primarily at maker rates and concentrate activity near 50-cent levels, face effective fee rates of 0.44 cents per side -- roughly 25% of the taker rate. This differential creates a structural cost advantage for professional participants.

2.4 Market Microstructure

Analysis of Becker's (2026) 72.1 million trade records reveals a pronounced structural asymmetry. The NO side of the

order book carries approximately 40x more resting depth than the YES side across all strikes and time horizons. Market makers earn an average of +1.12% per contract while takers lose -1.12%, consistent with the classical market-making model where professional liquidity providers sell insurance to retail directional flow. This asymmetry is persistent across the full 385-day sample and does not vary meaningfully with volatility regime or time of day.

A representative order book snapshot for a BTC 5-minute contract at $95,100 with 3 minutes remaining illustrates the typical depth imbalance:

Side	Level 1	Level 2	Level 3	Total
YES bids	8 @ 42c	15 @ 41c	6 @ 40c	29 contracts
NO bids	185 @ 55c	240 @ 54c	310 @ 53c	735 contracts

The 25:1 ratio (29 vs. 735) is representative; the dataset average is approximately 40:1. This asymmetry reflects the market's core dynamic: algorithmic market makers on the NO side absorbing directional YES-side flow from retail participants. The implication for model-based trading is significant: a trader taking the YES side faces thin liquidity and wide spreads, while taking the NO side means competing with professional market makers who have infrastructure, latency, and cost advantages.

3. Data

3.1 Becker (2026) Dataset

The primary dataset is Becker (2026), which provides trade-level records for 877,606 settled Kalshi crypto range contracts spanning 385 continuous trading days from March 2025 through March 2026. The dataset includes 72.1 million individual trades with timestamps (millisecond precision), prices (cent-level), quantities, and side indicators (buy/sell, maker/taker), enabling both aggregate analysis and microstructure research. To our knowledge, this is the largest publicly available trade-level dataset for any CFTC-regulated prediction market.

The dataset covers five crypto assets: BTC, ETH, DOGE, XRP, and SOL, across three time horizons: five-minute, hourly, and daily contracts. Settlement outcomes are included for all 877,606 contracts, allowing direct measurement of pricing accuracy. The asset breakdown is heavily concentrated: BTC accounts for 62% of contracts and 71% of volume, ETH for 28% and 22%, with DOGE, XRP, and SOL sharing the remaining 10% of contracts and 7% of volume.

3.2 Signal Extraction

From the full dataset, we extracted 28,496 qualifying signals -- instances where our model identified a mispricing large enough to potentially warrant a trade. The qualification criteria required three conditions simultaneously: (1) model fair value diverged from market price by more than the expected taker fee for that price level, creating a positive expected value after costs; (2) the contract had at least one executed trade within the prior 60 seconds, proving active quoting and providing a reliable reference price; and (3) sufficient time remained before settlement (at least 2 minutes) for the pricing model to be meaningful, as sub-minute pricing is dominated by the TWAP mechanism rather than our model's volatility estimates.

The asset-level breakdown reveals substantial heterogeneity in both signal frequency and edge magnitude:

Asset	Signals	Avg P&L	Win Rate
BTC	18,865	+0.4c	68.6%
ETH	9,612	+2.6c	42.5%
DOGE	696	+1.7c	50.7%
XRP	44	+9.9c	59.1%

ETH drove the majority of the aggregate edge despite having roughly one-third fewer signals than BTC. The average ETH signal produced +2.6 cents versus +0.4 cents for BTC, suggesting either greater mispricing in ETH markets or better model fit to ETH's volatility characteristics. BTC signals were plentiful but thin, consistent with a more efficiently priced market where professional market makers concentrate their attention. XRP and DOGE had too few signals for statistical confidence, though the XRP average of +9.9 cents hints at possible inefficiency in thinly traded assets.

3.3 Calibration Data

Probability calibration used 5.48 million contract observations spanning the full 385-day sample. Each observation consists of a model-predicted probability and a binary settlement outcome (1 if the contract settled YES, 0 otherwise), enabling direct measurement of calibration accuracy via Brier scores and reliability diagrams. The calibration dataset was constructed by sampling every active contract at 30-second intervals, producing approximately 14,200 observations per contract on average. This dense temporal sampling ensures that the calibration captures the full lifecycle of each contract, from listing to settlement, with sufficient granularity to detect time-varying biases.

3.4 Kou Model Calibration Data

For the Kou jump-diffusion model, we calibrated on 5,152 hourly BTC price returns extracted from the same 385-day period. Returns were computed as log-price differences between consecutive hourly closes, excluding periods with missing data (exchange outages, API gaps). The hourly return distribution showed kurtosis of 32.9 -- a figure that requires context. A normal distribution has kurtosis of 3.0. U.S. equity returns typically show kurtosis of 5-10. Commodity futures exhibit kurtosis of 8-15. The BTC value of 32.9 indicates extreme fat tails, with 3-sigma moves occurring approximately every 2 days versus every 15 days under normality -- an 8x elevation in tail frequency. At the 4-sigma level, the discrepancy is even more extreme: observed frequency is approximately 110x higher than the normal prediction.

The distributional characteristics of the hourly returns motivate the Kou model's double-exponential jump component: the excess kurtosis cannot be captured by adjusting diffusion volatility alone. A skewness of -0.3 indicates slight negative asymmetry, consistent with the well-documented tendency for crypto drawdowns to be sharper than rallies, though the asymmetry is modest compared to the kurtosis effect. The practical implication for pricing is significant: a log-normal model calibrated to match the central 95% of the return distribution will dramatically underestimate the frequency of returns in the tails, precisely the region that determines the prices of out-of-the-money binary contracts. For a contract 3 strikes from the money with 30 minutes remaining, the tail probability matters more than the central tendency, and a model that underestimates tail risk by a factor of 8-110x will systematically misprice these contracts.

4. Model Development

4.1 Black-Scholes Adaptation

The standard Black-Scholes framework prices a European binary call option paying $1 if the underlying asset price exceeds the strike price at expiry, and $0 otherwise. The fair value of this contract is given by the risk-neutral probability of expiring in-the-money:

$$Fair\ value = N(d?)$$

where d? = [ln(S/K) - 0.5sigma2T] / (sigmasqrtT), S is the current spot price, K is the strike price, sigma is the annualized implied volatility, T is the time to expiry in years, and N(·) is the standard normal cumulative distribution function. For Kalshi crypto range contracts, the adaptation requires three modifications: converting from European call options to range brackets, incorporating the TWAP settlement mechanism, and calibrating to the specific volatility characteristics of five-minute and hourly crypto returns.

We adapted this framework for Kalshi's specific settlement mechanism. The TWAP variance reduction factor of 0.554, derived in Section 2.2, is applied as a volatility multiplier that transitions linearly with time to expiry. The effective volatility is defined as:

$$\sigma_{effective}(\tau) = \sigma \times [1 - (1 - \sqrt{0.554}) \times \max(0, 1 - \tau/5)]$$

where ? is minutes to expiry. For ? < 1 minute (full TWAP regime), sigma_effective ~ sigma x 0.744. For ? > 5 minutes, sigma_effective = sigma (no adjustment). The linear transition reflects the increasing relevance of TWAP mechanics as settlement approaches. The transition window of 1-5 minutes was selected based on empirical analysis of pricing errors as a function of time-to-expiry; shorter windows produced discontinuities in the pricing surface, while longer windows over-adjusted contracts that were still primarily driven by spot price dynamics.

The initial implementation produced a backtest return of +221% on a $5,000 starting bankroll over 385 days -- a result that should have raised immediate suspicion. The annualized Sharpe ratio of 8.1, if genuine, would represent the most profitable systematic trading strategy ever documented in the academic literature. As we detail in Section 6, this extraordinary performance was entirely artifactual, the product of three simultaneous implementation bugs that were masked by our calibration procedure.

4.2 Isotonic Calibration

We applied isotonic regression to map model-predicted probabilities to empirically realized frequencies. Isotonic regression fits a monotonically non-decreasing function to the (prediction, outcome) pairs, providing a nonparametric calibration that preserves rank ordering while correcting systematic biases. The method is widely used in probabilistic forecasting (weather, medical diagnostics, credit scoring) and is considered a standard post-processing step for any probability model. Its appeal lies in its minimal assumptions: the only requirement is that higher predictions should correspond to higher observed frequencies, which is a necessary condition for any useful probability model.

The in-sample Brier score was 0.0103 across 5.48 million observations, improving to 0.0099 post-calibration. The modest improvement (4%) suggested the model was already well-calibrated -- a conclusion that, as we discuss in Section 6, was dangerously misleading. The calibration step was doing far more than minor refinement; it was compensating for fundamental model errors and producing the appearance of accuracy from a structurally broken model.

The calibration step is standard in probabilistic modeling and generally considered benign. Our central methodological finding is that it can mask fundamental model errors, creating a false sense of validation. The mechanism is subtle: isotonic regression is a flexible nonparametric estimator that can approximate any monotonic function. If the model's errors are approximately monotonic in the predicted probability -- which is common for systematic biases like sign errors or scaling errors -- isotonic regression will silently correct them. The corrected model will appear well-calibrated in sample, pass Brier score checks, and produce plausible backtests, while the underlying model remains fundamentally wrong.

4.3 Probability Cap Model

Our second model variant applies a hard cap at P = 0.45 for the maximum predicted probability of any single bracket. The motivation is empirical: analysis of the calibration dataset revealed that realized bracket frequencies plateau at approximately 50% regardless of model confidence above 45%. This pattern is consistent with the extreme kurtosis of crypto returns (32.9 as measured in our sample), which produces heavier tails than the log-normal assumption implies. The excess tail mass pulls probability away from the most likely bracket toward the tails, creating a natural ceiling on the probability of any single bracket that is lower than the Gaussian model predicts.

We observed the same overconfidence pattern in our weather prediction market analysis (OddsReference Research, 2026c): when the NWS ensemble model predicted 50%+ probability for a temperature bracket, the bracket was hit only 44.3% of the time. The cross-domain parallel -- crypto and weather, fundamentally different phenomena -- suggests this

is a structural property of Gaussian-based pricing models applied to bounded-outcome events, not a domain-specific artifact. Any model that uses a thin-tailed distribution to price a thick-tailed process will overestimate the probability of the most likely outcome and underestimate the tails.

The cap implementation is straightforward: any model-predicted bracket probability exceeding 0.45 is clipped to 0.45, and the excess probability is redistributed proportionally to the remaining brackets in the strike ladder. The redistribution preserves the total probability mass at 1.0 and maintains the relative ordering of non-capped brackets. This is mathematically equivalent to applying a soft maximum function to the bracket probability distribution.

4.4 Kou Jump-Diffusion Model

The Kou (2002) double-exponential jump-diffusion model extends geometric Brownian motion with a compound Poisson jump component. The log-price process follows:

$$dS/S = ?dt + sigmadW + dJ$$

where W is standard Brownian motion and J is a compound Poisson process with intensity ?. Jump sizes are double-exponentially distributed, allowing for asymmetric upward and downward jumps with different decay rates. Specifically, the log-jump size Y has probability density:

$$f(y) = p \cdot ?? e^{-?? y} \cdot 1_{y>=0} + q \cdot ?? e^{?? y} \cdot 1_{y<0}$$

where p + q = 1, p is the probability of an upward jump, ?? > 1 controls the decay rate of upward jumps, and ?? > 0 controls the decay rate of downward jumps. The constraint ?? > 1 ensures finite expected asset price. This parameterization has the analytical convenience that the moment-generating function of the double-exponential distribution has a closed form, enabling efficient computation of option prices via Fourier inversion.

We calibrated the model on 5,152 hourly BTC returns using maximum likelihood estimation. The calibrated parameters are:

Parameter	Symbol	Value	Interpretation
Jump intensity	?	2.3 / day	~1 jump per 10.4 hours
Excess kurtosis	?	32.9	vs. 3.0 (normal), 5-10 (equities)
3-sigma frequency	--	Every 48 hours	vs. 370 hours under normality
4-sigma frequency	--	Every 12 days	vs. 3.6 years under normality

At a typical hourly volatility of 0.4%, a 4-sigma move represents a 1.6% price swing -- approximately $1,500 at a $95,000 BTC price. The Kou model captures these events explicitly through the jump component, rather than relying on the log-normal tail (which underestimates their frequency by a factor of approximately 110 at the 4-sigma level). For pricing five-minute and hourly binary options, this improved tail modeling is potentially material: a single large jump during the contract window can move the settlement price across multiple bracket boundaries, and the model that better captures this possibility should produce more accurate prices.

The practical question is whether this improved tail modeling translates to better pricing in the Kalshi context. The Kou model requires calibrating 5 parameters (?, sigma, ?, ??, ??) versus 1 for Black-Scholes (sigma), substantially increasing the risk of overfitting. Moreover, the TWAP settlement mechanism dampens the very tail events that the Kou model is designed to capture: a large price spike during the contract window may not affect the trimmed mean settlement if it occurs as a brief outlier that falls within the 20% trimmed samples. The interaction between jump dynamics and TWAP settlement is complex and not amenable to closed-form analysis, requiring Monte Carlo simulation for accurate pricing.

5. Initial Results

The initial backtest across 28,496 qualifying signals produced results that appeared extraordinary. The following table summarizes the key performance metrics:

Metric	Value
Average edge per signal	+6.7 cents
Simulated return (385 days, $5,000 start)	+221%
Daily Sharpe ratio	8.1
Win rate	74.2%
Maximum drawdown	12.3%
Profit factor	3.8

Walk-forward validation, splitting the dataset into training (first 250 days) and testing (remaining 135 days) periods, showed +6.2 cents per signal in the test period -- apparently robust to regime change. The walk-forward Sharpe declined modestly from 8.4 (training) to 7.6 (test), which appeared consistent with normal parameter uncertainty rather than overfitting. The calibration Brier score of 0.0099 further reinforced confidence in the model's accuracy.

These results should have raised alarm. A Sharpe ratio of 8.1 on a systematic strategy without human discretion is implausible in any liquid market. For comparison, Renaissance Technologies' Medallion Fund -- widely considered the most successful systematic trading strategy in history -- achieves estimated Sharpe ratios of 2-3 before leverage. A Sharpe of 8.1 implies either a fundamental market inefficiency of historic proportions, an error in the data, or an error in the model. The academic literature on prediction market efficiency (Wolfers and Zitzewitz, 2004; Snowberg and Wolfers, 2010) finds typical mispricings on the order of 1-3 percentage points, far too small to support the 6.7-cent average edge our model claimed to identify.

We chose to validate through paper trading -- executing the model's signals in real time without risking capital. Paper trading provides a definitive out-of-sample test: the model sees live market data it has never been trained on, and the execution environment imposes real-world constraints (latency, slippage, liquidity) that backtests cannot capture. The paper trading period ran for 37 days beginning in late January 2026. The results were immediate and unambiguous: -29.5% cumulative return with zero positive weeks. The model systematically entered positions that immediately moved against it, with the largest losses concentrated in BTC contracts priced between 85 and 95 cents.

The 250 percentage point discrepancy between backtested (+221%) and paper-traded (-29.5%) performance could not be explained by regime change, volatility shifts, or fee miscalculation. The discrepancy was too large and too systematic. Something was structurally wrong with the model. We spent 14 days systematically debugging, which led to the discovery of three simultaneous implementation errors documented in the following section.

6. Bug Discovery and Correction

This section documents the methodological core of the paper. The discovery process illustrates how standard calibration techniques can create false confidence in broken models, and provides concrete diagnostic lessons for researchers working with calibrated probabilistic forecasts.

6.1 Paper Trading Diagnostics

The initial diagnostic was a pattern of 11 consecutive losses on BTC contracts in the 85-95 cent probability range. This cluster was statistically implausible under the model's predicted win rate of 74.2%: the probability of 11 consecutive losses at that win rate is $0.258^{11} = 0.0000036$, or roughly 1 in 280,000. Hand calculation of model fair values for these contracts revealed 40-60 percentage point disagreements with market prices -- far too large to explain by volatility misestimation or calibration drift. The disagreements were structural: the model was computing fundamentally incorrect probabilities for a systematic subset of contracts.

The pattern that emerged was directional: the model overpriced contracts where spot was above strike and underpriced contracts where spot was below strike. This inversion pattern was the first clue pointing toward a sign error in the core pricing formula.

6.2 Bug 1: Subtraction Order

The log-moneyness term in the d? computation used strike minus spot instead of spot minus strike:

$$\textit{Incorrect: } \ln(K/S) = \ln(95000/95500) = -0.0053$$

$$\textit{Correct: } \ln(S/K) = \ln(95500/95000) = +0.0053$$

This sign error inverted the model's probability estimates for in-the-money contracts. A contract that should have been priced at 65% (in-the-money, spot above strike) was priced at approximately 35% (out-of-the-money). The error affected approximately 40% of contracts -- those where spot was above strike at the time of signal generation. For contracts where spot was below strike, the sign error was inconsequential because both formulations agree on the direction of the log-moneyness term (negative in both cases, though with different magnitudes).

The critical observation: isotonic calibration compensated for this error. In-sample, the calibration routine learned a mapping that effectively re-inverted the broken probabilities. The historical data showed that when the model said 35%, the contract settled YES approximately 65% of the time, and isotonic regression dutifully fit this relationship. The calibrated model produced correct-looking probabilities despite the underlying model being directionally wrong for 40% of contracts.

In paper trading, the calibration table -- fitted on historical data with different volatility characteristics, different strike distributions, and a different market-making regime -- failed to compensate. The model's systematic sign errors were exposed as direct trading losses. The calibration mapping from the training period assumed a specific statistical relationship between model error and predicted probability; when that relationship shifted even modestly, the compensating correction broke down and the raw model error dominated.

This is the most dangerous aspect of calibration as a validation tool. A sufficiently flexible calibration method (and isotonic regression is highly flexible, capable of approximating any monotonic function on the unit interval) can compensate for any monotonic transformation of the input -- including a sign reversal. The calibrated model appears correct, the Brier score is good, the backtest returns are positive, but the model is fundamentally broken and will fail out-of-sample when the statistical relationship between model error and market conditions changes.

6.3 Bug 2: Time-to-Expiry Units

The model's volatility calculation expected time-to-expiry in minutes. The implementation provided hours. For a contract 30 minutes from expiry:

$$\textit{Incorrect: } \sigma\sqrt{T} = \sigma\sqrt{30} \text{ (treating 30 as hours = 1,800 minutes)}$$

$$\textit{Correct: } \sigma\sqrt{T} = \sigma\sqrt{30/60} \text{ (30 minutes = 0.5 hours)}$$

$$\textit{Error factor: } \sqrt{30 \times 60 / 30} = \sqrt{60} = 7.74\text{x}$$

This 7.7x overestimate of effective volatility caused the model to dramatically overestimate the probability of large price moves. Contracts 3 strikes away from the money with 20 minutes remaining -- which should have been priced near zero -- were modeled at 15-20 cents. The model was buying heavily into far out-of-the-money contracts in the final hour, losing on nearly all of them because the implied moves were approximately 60x too large in variance terms.

The discovery came from analyzing the temporal pattern of losses. Losses concentrated overwhelmingly in the final 60 minutes before settlement, with a sharp spike in the last 30 minutes. This pattern was inconsistent with a model that correctly estimated volatility: if the model were unbiased, loss frequency should be approximately uniform across time-to-expiry bins. The concentration in the final hour pointed to a time-dependent error, and the magnitude of the error

-- 7.7x at 30 minutes -- was exactly consistent with a minutes-versus-hours unit confusion.

The units bug interacted with the sign bug in a way that amplified both errors. The inflated volatility spread probability mass across a wider range of strikes, creating more apparent signals at extreme prices. Combined with the sign inversion, which redirected model attention to the wrong side of the strike distribution, the two bugs together produced a pattern of confidently wrong signals at extreme prices -- exactly the high-conviction, high-loss pattern we observed in paper trading.

6.4 Bug 3: Bracket Interpretation

Kalshi's API returns a b_value field for each range contract. We interpreted this as the bracket floor (lower boundary of the price range). It is actually the bracket center.

For BTC contracts with $100 bracket width, a b_value of 95,000 means the bracket spans $94,950 to $95,050, not $95,000 to $95,100. The $50 offset is small in absolute terms -- 0.05% of a $95,000 BTC price -- but shifts probabilities by 0.5-3 percentage points for near-the-money contracts. For a contract at the money, where the probability surface is steepest, a $50 shift can move the fair value from 48% to 51%, changing the model's trade direction from sell to buy.

Verification was direct and conclusive: 14 contracts in our test set settled at prices that fell outside our assumed bracket boundaries but inside the correct (center-based) boundaries. All 14 anomalies -- cases where the model predicted YES settlement but the price was apparently outside the bracket -- were resolved by the center interpretation. The b_value interpretation was subsequently confirmed by direct communication with the Kalshi API documentation and cross-validation against published settlement results.

6.5 Combined Impact

The following table summarizes the combined impact of all three bug fixes on model performance:

Metric	Before Fixes	After Fixes
Backtest return	+221%	+7.0%
Avg edge per signal	+6.7c	+1.2c
Win rate	74.2%	59.6%
Paper trading return	-29.5%	Not retested
Sharpe ratio	8.1	4.5
Max drawdown	12.3%	8.7%

The corrected model shows a fundamentally different profile. The edge drops from +6.7 to +1.2 cents per signal -- a reduction of 82%. The win rate falls from 74.2% to 59.6%, moving from an implausible level to one consistent with a modestly predictive model operating in an approximately efficient market. The Sharpe ratio halves from 8.1 to 4.5, which is still high but no longer outside the range of published systematic strategies. The +221% return becomes +7.0%, barely exceeding the risk-free rate and well within the range of noise for a strategy with this level of turnover.

6.6 The Calibration Trap

The central methodological finding deserves explicit treatment, as it has implications extending well beyond crypto markets to any field that uses calibrated probabilistic forecasts.

Isotonic regression is a monotonic nonparametric estimator -- it maps input probabilities to output probabilities while preserving rank order. If the input probabilities are systematically biased (as ours were, due to three simultaneous bugs), isotonic regression will learn to correct for that bias using historical data. The correction is sample-specific: it fits the exact pattern of bias present in the training data. When conditions change -- different volatility regime, different strike

distribution, different market microstructure -- the correction fails, and the underlying model error is exposed.

The danger is twofold. First, the corrected model appears well-calibrated in-sample, with Brier scores comparable to (or better than) a correctly specified model. Our broken model achieved a post-calibration Brier score of 0.0099, which would be considered excellent by any standard metric. Second, the corrections are sample-specific: they compensate for the exact bias pattern in the training data and fail when conditions change. Walk-forward validation, which is often considered a robust out-of-sample test, did not catch the error because the bias pattern was sufficiently stable over the 385-day sample for the calibration to generalize between the training and test periods.

A diagnostic principle emerges: if removing calibration from a model dramatically changes its outputs, the underlying model is likely misspecified. In our case, the raw (uncalibrated) model after bug fixes produced well-calibrated probabilities with a Brier score of 0.0103 across 5.48 million observations. Calibration improved this marginally to 0.0099 -- a 4% improvement. Before bug fixes, calibration improved the Brier score by 38% -- from 0.0159 to 0.0099. The 38% improvement should have been a red flag: the calibration was doing too much work, indicating that the underlying model was producing systematically wrong probabilities that required substantial correction.

We propose a concrete diagnostic: compute the ratio of post-calibration to pre-calibration Brier score. If the ratio is below 0.80 (i.e., calibration improves the score by more than 20%), investigate the underlying model for misspecification before proceeding with backtests or trading. This threshold is based on our experience -- our correctly specified model showed a 4% improvement, our broken model showed 38% -- and should be treated as a heuristic rather than a formal statistical test.

7. Corrected Results

7.1 Black-Scholes (Clean)

The corrected Black-Scholes model with TWAP adjustment was re-evaluated across the full 385-day sample. The model produced 28,484 qualifying signals after applying the corrected signal extraction criteria (the slight reduction from 28,496 reflects edge cases where the corrected bracket boundaries moved contracts across the qualification threshold).

Metric	Value
Qualifying signals	28,484
Win rate	59.8%
Average taker P&L	+1.15 cents
Average maker P&L	+1.80 cents
Average zero-fee P&L	+2.80 cents
Total taker P&L	+$328
Total maker P&L	+$513
Sharpe ratio (daily)	4.5
Maximum drawdown	8.7%

The corrected model retains genuine predictive power -- the +1.15 cents average taker P&L is statistically significant with a t-statistic of 106.47 -- but the economic magnitude is insufficient to support a viable trading strategy. The total taker P&L of $328 over 385 days represents a 6.6% return on a $5,000 bankroll, below the contemporaneous risk-free rate of approximately 4.5% annualized. At maker fees, the picture improves modestly to $513 total P&L (10.3% return), but achieving consistent maker fills requires infrastructure and market-making capabilities that most individual traders lack.

The zero-fee P&L of +2.80 cents per signal provides a useful decomposition of the market's efficiency. The model identifies genuine mispricings of approximately 2.8 cents on average, but the transaction cost envelope of 1.58 cents

(taker fee) plus 1.5 cents (entry slippage) consumes the entire edge. The market is efficiently priced not because prices are perfectly accurate, but because the cost of exploiting inaccuracies exceeds the inaccuracies themselves.

7.2 Capped Black-Scholes

The probability-capped model, which clips maximum bracket probability at 0.45, produced nearly identical results to the uncapped version:

Metric	Uncapped	Capped (P <= 0.45)
Signals	28,484	28,484
Win rate	59.8%	59.6%
Avg taker P&L	+1.15c	+1.20c
Total taker P&L	+$328	+$350
Avg maker P&L	+1.80c	+1.80c

The cap barely changes results because most actionable signals already fall in the 0-40% probability range where the cap does not bind. The overconfident 50%+ zone generates few signals since market prices are already close to model prices there. This is consistent with efficient pricing at the money: the market accurately prices high-probability brackets, and mispricings concentrate in the tails where both model uncertainty and market thinness are greatest.

The marginal improvement from +1.15 to +1.20 cents per signal is not statistically significant ($p = 0.34$ for the paired difference), confirming that the cap does not meaningfully improve model performance. The overconfidence problem, while real in aggregate calibration statistics, does not manifest as a tradeable mispricing because the market already incorporates the same adjustment implicitly through its pricing of at-the-money contracts.

7.3 Kou Jump-Diffusion

The Kou model, with its explicit jump component calibrated on 5,152 hourly BTC returns, produced marginally better results than the Black-Scholes variants:

Metric	Black-Scholes	Kou
Signals	28,484	28,684
Win rate	59.8%	57.4%
Avg taker P&L	+1.15c	+1.38c
Total taker P&L	+$328	+$395
Avg maker P&L	+1.80c	+1.96c
Avg zero-fee P&L	+2.80c	+2.96c

Kou produces marginally better results: +1.38 cents versus +1.15 cents per signal, +$395 versus $328 total P&L. The improvement of +0.23 cents per signal ($67 total over 385 days) is statistically significant at the 5% level (paired t-test, $p = 0.018$) but economically negligible. The $67 annual improvement is the reward for calibrating 5 parameters instead of 1, implementing Monte Carlo simulation for TWAP-adjusted pricing, and accepting the higher overfitting risk inherent in the richer model. The complexity-adjusted value proposition is poor.

At zero fees, both models converge: +2.80 cents (Black-Scholes) versus +2.96 cents (Kou). The market misprices contracts by roughly 3 cents on average; fees consume most of this edge regardless of model sophistication. This convergence suggests that the binding constraint on profitability is not model accuracy but execution cost -- a finding consistent with the broader market microstructure literature, which emphasizes that statistical edge and tradeable edge are distinct concepts separated by the transaction cost envelope.

7.4 Statistical Significance

The corrected +1.2 cent edge is highly statistically significant: 95% confidence interval [+1.15, +1.25], t-statistic 106.47, $p < 0.0001$. The large t-statistic reflects the high signal count (28,484) rather than a large effect size. In standardized terms, the edge is approximately 0.63 standard deviations -- a modest effect that achieves significance through sample size rather than magnitude.

The edge is real in a statistical sense -- the model identifies genuine mispricings with high confidence. It is simply too small to survive transaction costs and execution constraints. This distinction between statistical and economic significance is critical: a researcher testing only for statistical significance would conclude that the model identifies profitable opportunities, while a practitioner incorporating execution costs would conclude the opposite.

7.5 The Directional Accuracy Paradox

One of the most counterintuitive findings from the corrected model: the model predicts the direction of contract settlement with 89.2% accuracy on far out-of-the-money contracts (priced below 10 cents), yet loses money on those contracts. The explanation involves the fundamental payoff asymmetry of binary options.

For a 5-cent contract (5% implied probability of settling YES), correctly predicting NO earns 5 cents 89.2% of the time, while incorrectly predicting NO (the event occurs despite low probability) costs 95 cents 10.8% of the time:

$$E[P\&L] = 0.892 \times \$0.05 - 0.108 \times \$0.95 = \$0.0446 - \$0.1026 = -\$0.058$$

High accuracy, negative expected profit. This result illustrates that in binary options, accuracy and profitability are fundamentally decoupled. What matters is the calibration between confidence and realized probability -- the mapping from model-predicted P to empirical settlement frequency -- not the raw directional hit rate. A model that correctly predicts 89% of 5-cent contracts is actually underperforming: if the true probability is 5%, a correct model should predict NO 95% of the time, not 89.2%. The 5.8 percentage point underperformance (89.2% vs. 95.0%) represents a systematic overestimation of tail probabilities that generates losses despite high apparent accuracy.

This paradox has practical implications for model evaluation. Researchers accustomed to evaluating classifiers by accuracy or area under the ROC curve may be misled by high accuracy scores that coexist with negative expected P&L. In binary options markets, the proper evaluation metric is the Brier score or, more directly, the average P&L per signal after accounting for the price at which the trade is executed.

8. Execution Constraints

8.1 Exit Liquidity

Of the 28,484 contracts where the corrected model identified tradeable edge, 68% showed zero exit liquidity -- no resting orders available on the opposite side for an immediate sell. The only exit for these positions is binary settlement at $1.00 or $0.00. Trade-out strategies, which are standard in traditional options markets for managing risk and locking in profits, are therefore not viable for the majority of positions.

We simulated trade-out strategies at various capture rates -- the fraction of model-identified edge that a trader captures by exiting before settlement:

Capture Rate	Final Value ($5K start)	Return	Max Drawdown
10%	$7.42	-99.9%	99.9%
20%	$9.46	-99.8%	99.8%
40%	-$13.81	-100.3%	100.3%
Hold to expiry	$5,328	+6.6%	8.7%

The near-total losses at partial capture rates result from adverse selection. At a 10% capture rate, the positions the trader exits are the easy ones -- they moved in the trader's favor quickly, indicating the market agreed with the model's assessment. The remaining 90% are positions where the market disagrees with the model, and these positions tend to be wrong. The trader systematically sells winners and holds losers, a destructive pattern that compounds into near-total capital loss. Only the hold-to-expiry strategy, which avoids the adverse selection problem entirely by never trading out, preserves capital.

8.2 Volume-Edge Correlation

The correlation between model-identified edge (absolute difference between model fair value and market price) and contract trading volume is 0.849. This strong positive correlation means that the largest mispricings exist in the least liquid contracts. This is expected from a market efficiency perspective -- efficiency is a function of attention, capital allocation, and competitive market-making -- but it imposes a hard capacity constraint on model-based trading. The contracts where the model finds the most edge are precisely those where execution is most difficult: wide spreads, thin depth, and high slippage.

Decomposing the correlation by asset class reveals that the relationship is strongest for DOGE and XRP ($r = 0.91$ and 0.93 respectively), where market-making activity is sporadic and mispricings can persist for extended periods. For BTC, the correlation is lower ($r = 0.78$) but still substantial, indicating that even in the most liquid crypto binary market, the edge-liquidity tradeoff is binding. This inverse relationship between edge magnitude and execution feasibility is a common feature of financial markets, often described as the limits of arbitrage (Shleifer and Vishny, 1997). The specific form it takes in crypto binary options -- where the constraint is not capital or margin requirements but simply the absence of counterparties -- is distinctive and reflects the market's concentrated liquidity structure.

8.3 Capacity Ceiling

Based on observed fill rates of 3.6-8.8% of signals and entry slippage of 1.5-1.6 cents, the capacity ceiling for a single model-based trader is \$5,000-\$25,000 in deployed capital. Above this threshold, the trader's own order flow begins moving the market against the intended entry price, and the 1.2-cent edge disappears into slippage. The capacity estimate is based on three converging analyses: observed depth at the best price across 28,484 signal events, the historical fill rates for limit orders at model-determined prices, and the price impact function estimated from Becker's (2026) trade data.

At the upper bound (\$25,000 deployment with 8.8% fill rate), a trader would execute approximately 2,508 trades per year (28,484 signals x 8.8% fill rate). At +1.2 cents per trade, annual gross profit would be approximately \$30 -- well below the opportunity cost of capital, the infrastructure cost of running a real-time pricing model, or the market data fees required for live implied volatility data. The capacity constraint effectively renders the statistical edge untradeable at any meaningful scale.

8.4 Slippage Budget

The complete cost decomposition for a round-trip trade (entry and settlement) is as follows:

Component	Cost (cents)
Taker fee (average)	1.58
Entry slippage	1.50
Exit (settlement)	0.00
Total round-trip	3.08

Against a gross edge of 2.8-3.0 cents (zero-fee model P&L), the 3.08-cent execution cost leaves essentially nothing. The

market is efficient in the economic sense: the cost of exploiting identified mispricings equals or exceeds the mispricings themselves. At maker fees, the round-trip cost drops to approximately 2.0 cents (0.40 cents maker fee + 1.5 cents slippage + 0.0 cents settlement), leaving approximately 0.8-1.0 cents of net edge. However, achieving consistent maker fills requires sophisticated infrastructure: resting orders across 188 strikes with real-time requoting as the underlying price moves, latency-optimized execution, and continuous position management. This infrastructure cost further erodes the marginal profitability of maker-based strategies.

The hold-to-expiry constraint -- zero exit cost but no ability to trade out -- creates a binary risk profile that is unusual relative to traditional options trading. Each position either earns the full contract payout ($1.00) or loses the entire entry cost. There is no ability to cut losses on positions moving against the model, and no ability to take profits on positions moving in favor. The model must be right on average, with sufficient margin to cover the cost of being wrong, and this margin (1.2 cents at taker rates) is razor-thin.

9. Discussion

9.1 Efficiency Finding

Our three models converge on the same conclusion: Kalshi crypto binary options are approximately efficiently priced. The maximum model edge of +1.4 cents per signal (Kou at taker fees) is statistically significant but economically insufficient after execution costs. The market incorporates publicly available pricing information -- implied volatility from centralized exchanges, time decay, TWAP adjustment mechanics, and recent price momentum -- into contract prices with sufficient accuracy that the remaining mispricing falls within the transaction cost envelope.

This finding is consistent with our concurrent analysis of Kalshi weather prediction markets (OddsReference Research, 2026c), where an NWS ensemble model achieved 33% bracket accuracy (versus a 10% random baseline) but no profitable trading strategies across 1,506 settled contracts. The cross-domain convergence -- crypto and weather, fundamentally different phenomena with entirely different pricing models and data sources -- suggests that prediction market efficiency is driven by structural factors rather than domain-specific expertise. These structural factors include: public information availability (both crypto prices and weather forecasts are freely accessible), rapid resolution cycles (hourly for crypto, daily for weather, preventing information from going stale), and active market-making (professional liquidity providers on both platforms continuously arbitrage mispricings).

The efficiency finding is nuanced. The market is not perfectly efficient -- our models detect genuine statistical mispricings with high confidence -- but the mispricings are small enough that the transaction cost envelope absorbs them. This is consistent with the Grossman-Stiglitz (1980) framework: markets cannot be perfectly informationally efficient because traders require compensation for the costs of acquiring and processing information. The roughly 3-cent average mispricing we observe can be interpreted as the equilibrium information rent that compensates market makers for their pricing infrastructure.

9.2 Methodological Contribution

The paper's primary methodological contribution is the documentation of a specific and dangerous failure mode in quantitative model validation. Three simultaneous implementation bugs -- individually subtle, collectively producing a 5.5x overestimate of model edge -- were masked by isotonic calibration and survived walk-forward validation. Only paper trading, which introduced genuinely out-of-sample conditions with a different volatility regime and market microstructure, exposed the errors.

This has implications for any research using calibration as a model validation step, which includes most probabilistic forecasting in weather, medicine, credit scoring, and quantitative finance. We propose a diagnostic principle: compute model performance with and without calibration. If calibration accounts for more than a marginal improvement (we

suggest >20% of the Brier score improvement), investigate the underlying model for misspecification before proceeding with backtests, trading, or production deployment.

The three bugs we discovered are individually common in quantitative modeling: subtraction order errors (equivalent to sign errors in physics), unit mismatches (equivalent to the Mars Climate Orbiter failure), and API field misinterpretation (equivalent to off-by-one errors in computer science). What makes our case study novel is the interaction between these common errors and the calibration step, which transformed three obvious bugs into a plausible-looking trading strategy that passed standard validation checks.

9.3 Comparison to Related Work

Becker (2026) documented the maker-taker dynamic in Kalshi crypto markets, finding +1.12% average maker returns versus -1.12% taker returns across 72.1 million trades. Our model-based analysis corroborates this from the opposite direction: model edge at taker rates (+1.2 cents) closely approximates the taker penalty identified by Becker, suggesting the market's efficiency is driven primarily by market maker activity. The convergence between Becker's aggregate microstructure analysis and our model-based efficiency test strengthens both findings.

Konczal (2025) argued that Kou jump-diffusion models should outperform Black-Scholes for BTC options pricing due to the extreme kurtosis of crypto returns. Our results partially support this claim -- Kou does produce marginally better edge (+0.23 cents per signal) -- but the improvement is too small to be practically meaningful and comes at substantial calibration complexity cost. The 5-parameter Kou model's marginal advantage over the 1-parameter Black-Scholes suggests that the binding constraint on profitability is execution cost, not model accuracy. Even a perfect pricing model would face the same 3.08-cent transaction cost, and the incremental improvement from better tail modeling (0.23 cents) is overwhelmed by this cost.

Our results contribute to the broader prediction market efficiency literature initiated by Wolfers and Zitzewitz (2004) and Manski (2006). While these foundational papers focused on political prediction markets with long time horizons, our analysis extends the efficiency finding to ultra-short-duration (5-minute and hourly) markets with continuous pricing. The efficiency finding holds across time horizons, suggesting that prediction market efficiency is a robust phenomenon driven by market structure rather than event-specific factors.

9.4 Implications for Market Participants

For retail traders, the message is clear: model-based edge exists but is thin, capacity-constrained, and insufficient to cover transaction costs. The average retail taker loses 1.12% per contract (Becker, 2026), and our analysis confirms that this loss is structural rather than correctable through better modeling. Retail participants in Kalshi crypto markets are, in aggregate, paying for entertainment value and directional exposure rather than earning risk-adjusted returns.

For platform design, the results suggest Kalshi's fee structure is set at approximately the right level to balance revenue with market efficiency. A substantially lower fee would make model-based trading profitable, potentially attracting more sophisticated algorithmic flow that could crowd out retail participation. A higher fee would widen spreads, reduce liquidity, and potentially drive activity to unregulated alternatives. The current fee level sits in the narrow band where the market functions efficiently while remaining attractive to both sides of the ecosystem.

For regulators, the finding that CFTC-regulated crypto prediction markets are efficiently priced, even at approximately 13 months old and with retail-dominated flow, suggests the regulatory framework is functioning as intended. The DCM structure provides sufficient transparency and competitive conditions for prices to reflect available information. The TWAP settlement mechanism provides meaningful manipulation resistance, and the position limits prevent outsized concentrated risk. These structural safeguards appear to be sufficient for maintaining market integrity in the novel domain of crypto binary options.

9.5 Limitations

Our analysis is subject to several important limitations. First, the study uses a single dataset (Becker, 2026), which, despite its comprehensiveness (72.1 million trades), represents only the Kalshi platform. Cross-platform validation using Polymarket or other crypto prediction markets would strengthen the efficiency finding but is complicated by different fee structures, settlement mechanisms, and regulatory environments.

Second, our analysis is limited to hold-to-expiry strategies. Market-making strategies -- which involve continuously quoting both sides of the order book and earning the bid-ask spread -- were not tested and may fare better. The +1.80 cents average maker P&L suggests that a well-executed market-making strategy could be profitable, but evaluating this requires modeling adverse selection risk, inventory management, and the latency competition with existing market makers.

Third, the capacity analysis is theoretical, based on observed fill rates and depth statistics rather than live execution data. A live trading test with actual capital would provide more definitive evidence on capacity constraints, slippage, and the interaction between model signals and market response.

Fourth, the Kou model calibration uses hourly returns, which may not fully capture intraday jump dynamics relevant to five-minute contracts. A sub-minute calibration sample might improve the model's performance on the shortest-duration contracts, though the additional data requirements and calibration complexity may not justify the incremental improvement.

10. Conclusion

We develop and test three pricing models for CFTC-regulated crypto binary options on Kalshi, using 877,606 settled contracts and 72.1 million trades from the Becker (2026) dataset. The models demonstrate genuine predictive power: the corrected Black-Scholes model achieves +1.2 cents average edge per signal ($t = 106.47$, $p < 0.0001$), and the Kou jump-diffusion model improves this to +1.4 cents. However, no model produces tradeable profits after taker fees (1.58 cents average), entry slippage (1.5 cents), and the binding constraint of 68% zero exit liquidity. The capacity ceiling of $5,000-$25,000 for a single model-based trader renders the statistical edge economically irrelevant.

The paper's cautionary tale -- a +221% backtest collapsing to +7% after three bug fixes -- underscores the fragility of quantitative model validation. Isotonic calibration masked three simultaneous implementation errors, each individually subtle, producing plausible returns from a fundamentally broken model. Walk-forward validation did not catch the errors. Only paper trading, which introduced genuinely out-of-sample conditions, exposed the fundamental misspecification. We propose a diagnostic: if removing calibration changes model performance by more than 20%, investigate the underlying model for misspecification before proceeding.

Our cross-domain comparison with weather prediction markets -- where an NWS ensemble model achieved 33% bracket accuracy but zero profitable strategies across 1,506 events -- reinforces the conclusion: CFTC-regulated prediction markets are approximately efficient when public information is available and events resolve quickly. The convergence across crypto and weather domains, using entirely different models, data sources, and pricing methodologies, suggests this efficiency is a structural property of well-regulated prediction markets rather than a domain-specific finding.

The value of these markets lies not in the trading profits they offer to individual participants, but in their transparency, information aggregation properties, and the price signals they provide to the broader market ecosystem. The finding that a 13-month-old retail-dominated market can achieve near-efficient pricing of complex financial instruments is itself a testament to the power of competitive market mechanisms, even in nascent market structures. The rapid convergence to efficiency is consistent with Arrow's (1963) foundational insight that markets can aggregate dispersed information efficiently under competitive conditions, extended here to a novel class of ultra-short-duration binary contracts on highly volatile underlying assets.

Future research should extend this analysis in three directions. First, cross-platform comparisons between Kalshi and Polymarket would test whether the efficiency finding is platform-specific or generalizable across different regulatory and fee structures. Second, market-making strategy evaluation -- testing whether the +1.80 cents average maker edge translates to profitable strategies after accounting for adverse selection, inventory risk, and latency costs -- would complement our hold-to-expiry analysis. Third, the calibration trap documented in Section 6 deserves systematic investigation across other domains where isotonic or Platt calibration is standard practice, including medical AI, credit scoring, and weather forecasting. The interaction between model misspecification and calibration flexibility may be a more pervasive problem than the current literature recognizes.

References

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. American Economic Review, 53(5), 941-973.

Becker, E. (2026). Trade-level analysis of Kalshi crypto binary options: Market microstructure and maker-taker dynamics. Working Paper.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. Journal of Political Economy, 81(3), 637-654.

CF Benchmarks. (2025). TWAP methodology: Crypto settlement index documentation. CF Benchmarks Ltd.

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. American Economic Review, 70(3), 393-408.

Kalshi, Inc. (2025). Platform rules, fee schedule, and contract specifications. Retrieved from kalshi.com.

Konczal, M. (2025). Jump-diffusion pricing for crypto binary options: Kou model calibration on BTC hourly returns. Quantitative Finance Working Paper.

Kou, S. G. (2002). A jump-diffusion model for option pricing. Management Science, 48(8), 1086-1101.

Manski, C. F. (2006). Interpreting the predictions of prediction markets. Economics Letters, 91(3), 425-429.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. Journal of Financial Economics, 3(1-2), 125-144.

OddsReference Research. (2026a). Inside Kalshi's crypto binary options: $60M/day in five-minute bets. OddsReference Working Paper.

OddsReference Research. (2026b). Hold-to-expiry edge in CFTC-regulated crypto binary options. OddsReference Working Paper.

OddsReference Research. (2026c). Forecast efficiency in CFTC-regulated weather prediction markets. OddsReference Working Paper.

Snowberg, E., & Wolfers, J. (2010). Explaining the favorite-longshot bias: Is it risk-love or misperceptions? Journal of Political Economy, 118(4), 723-746.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. Journal of Economic Perspectives, 18(2), 107-126.