Kalshi Crypto Trading Bot

Investor Whitepaper

Gabriel Kagan

April 2026

Contents

Executive Summary

This document describes an automated trading platform for **Kalshi**, the first CFTC-regulated prediction market exchange in the United States. The system began with short-duration cryptocurrency contracts — binary options that settle every 15 minutes — and now runs a layered live business: 15M crypto across four assets, decided contracts (high-conviction overlay), late-window momentum, weekend and overnight discount entries, near-expiry low-price entries, and weather temperature NO-side. Adjacent product engines (S&P 500 intraday, hourly crypto, sports comebacks) collect calibration data in observation mode, with hourly currently kill-switched off after a March incident.

The platform monitors real-time data from multiple sources per vertical, estimates outcome probabilities using domain-specific models, and executes trades only when it identifies a clear edge over the market price. Every aspect of the strategy — from market selection to position sizing to execution — is designed around disciplined risk management and profit maximization.

Live trading results (auto-updated; last refresh 2026-05-02T12:18:34Z):

- **3,310** settled trades with a **92.6%** win rate (3,066W / 242L / 2 breakeven)
- Live trading with real capital since February 22, 2026
- Fully automated, always-on operation with complete audit trail
- Adjacent verticals — each follows the same shadow→validate→promote pipeline:
  - **Weather temperature** — 19 US cities; NO-side LIVE in 1-contract verification mode since Apr 11, YES-side observation
  - **S&P 500 intraday** — equity-adapted volatility model with VIX integration; observation only
  - **Live sports** — 28 leagues including NFL, NBA, EPL, ATP/WTA Tennis; observation only
  - **Hourly crypto** — disabled since Apr 18 after correlated multi-strike losses; env-gated re-enable path preserved

---

The Opportunity

Why Prediction Markets?

Prediction markets allow participants to trade contracts on the outcomes of real-world events. Unlike traditional financial markets, where pricing depends on complex fundamental analysis, prediction market contracts resolve to a simple binary outcome: the event either happens or it doesn't. This clarity creates a well-defined mathematical edge for quantitative approaches.

Why Kalshi?

Kalshi is the only CFTC-regulated prediction market exchange in the US, providing the legal and structural protections of a regulated financial venue. Its crypto contracts offer several attractive properties for systematic trading:

Property	Detail
Frequency	New contracts every 15 minutes, 24/7, across 4 cryptocurrencies
Duration	Each contract settles within 15 minutes — capital never locked up long
Binary outcomes	Pays full contract value if above threshold, nothing otherwise — no partial payoffs
Market depth	Hundreds of tradeable markets per day across multiple assets and strikes

The Systematic Edge

Retail traders on Kalshi typically check one price source and make intuitive judgments about whether a crypto asset will stay above a given level. This approach is slow, inconsistent, and unable to process the volume of markets available.

A systematic approach has structural advantages: it can process every market, every window, with consistent discipline — never skipping a promising trade due to fatigue, and never chasing a bad one due to emotion.

Strategy Overview

The bot follows a disciplined five-step process for every 15-minute trading window:

1. Observe

The system maintains real-time connections to multiple data sources simultaneously:

Source	Purpose
Coinbase + Kraken	Live cryptocurrency spot prices with sub-second updates from two exchanges
Deribit	Implied volatility data reflecting the market's forward-looking risk expectations
Kalshi	Current contract prices, orderbook depth, and real-time fill notifications via WebSocket

This multi-source approach means the bot sees price movements developing across global markets before they are reflected in Kalshi contract prices.

2. Estimate

For every active market, the bot estimates the probability that the underlying asset will stay above the contract's threshold for the remainder of the 15-minute window. This estimate incorporates:

- **Current price position** relative to the threshold
- **EGARCH-conditioned volatility** — how much the price is expected to move, accounting for clustering and leverage effects
- **Options-implied volatility** — what the derivatives market expects
- **Cross-exchange signals** — whether other exchanges are leading a move
- **Market-price blending** — 60% model / 40% market blend to prevent overconfidence

The probability model uses **Normal Inverse Gaussian (NIG) distributions** fitted specifically to each cryptocurrency's return characteristics. Unlike generic models, NIG captures both the heavy tails (large moves are more common than a bell curve predicts) and the asymmetry (upward and downward moves have different frequencies) unique to each asset.

3. Filter

Most markets are not worth trading. The system applies a rigorous multi-stage filter:

A market is rejected if any of the following apply:

- The model's estimated probability is too low (the contract is unlikely to pay out)
- The Kalshi price is too high (not enough profit potential) or too low (too much uncertainty)
- The edge after fees is insufficient — must exceed a price-dependent minimum (V-shaped: 0.25% at 80–90¢, dipping to 0.20% at 91–92¢, then climbing to 0.5% at 93–94¢, 0.75% at 95–96¢, and 1.0% at 97¢+) after worst-case taker fees
- The model and the market disagree by a suspicious margin
- Statistical inputs appear unreliable (extreme z-scores indicating potential data issues)

The vast majority of markets are correctly identified as unprofitable and filtered out — the system is highly selective.

4. Size

For the small number of markets that pass all filters, the bot determines the appropriate position size using an edge-tiered framework:

Fee-Adjusted Edge	Risk Fraction	Rationale
≥ 4.0%	25% of bankroll	Maximum conviction — strong statistical edge
≥ 2.5%	20%	High edge — data supports aggressive sizing
≥ 1.8%	15%	Solid edge — moderate allocation

Fee-Adjusted Edge	Risk Fraction	Rationale
≥ 1.2%	10%	Standard edge — controlled exposure
≥ 0.9%	7%	Lower edge — minimal sizing
≥ 0.7%	5%	Thin edge — small position
≥ 0.5%	3%	Marginal edge — small allocation
≥ 0.25%	2%	Minimum edge — smallest allocation

Position sizes are:

- **Automatically reduced during drawdowns** — below 85% of the rolling 7-day cash high-water mark, sizes halve; below 75%, they quarter; below 65%, trading halts entirely
- **Hard-capped** — no single trade exceeds 25% of bankroll regardless of model confidence; per-asset caps are tighter (BTC 15%, ETH 20%, SOL 15%, XRP 15%)

5. Execute

The bot uses a fee-minimizing execution strategy with intelligent escalation:

Execution Phase	Description
Maker-first	Places limit orders with `post_only` guarantees — no maker fee (taker fee only on escalation)
3-tier rejection handling	If maker rejected (locked spread): try degraded maker → taker IOC (re-verifying profitability at higher fees)
Direct taker below 180s	When <180s remain, skip maker entirely — data shows only 7.7% fill rate at low STC; direct taker strictly better
Real-time fill detection	Kalshi WebSocket provides instant fill notifications at zero API cost, with REST backup

Execution Phase	Description
Smart escalation	If unfilled: amend order in-place (faster than cancel + re-place), then fall back to IOC taker
Per-asset optimization	SOL bypasses maker entirely (direct taker) due to thin orderbooks; BTC uses shorter escalation wait (7s vs 15s default)
Near-certain overlay	Decided contract system (four live tiers: T1, T1B, T2, T2-Z25) identifies near-certain outcomes via extreme z-scores and routes to direct taker; T1/T1B/T2 sized 20%, T2-Z25 sized 10%. T2-Z2 returned to shadow Apr 22 after underperformance. 6 expansion shadows collect data for future tiers
Terminal momentum	Trades 96/98/99¢ contracts in the final 1–5 minutes with sizing capped to 25–500 contracts. Near-100% standalone WR
Low-price near-expiry	BTC 80–87¢ in the final 10–120 seconds, 50 contracts fixed, only with model conviction at the entry strike
Discount overlays	Weekend (Sat/Sun, 90¢+, STC≤600s) and overnight (weekday 04–11 UTC, 89¢+, STC≤600s) entries with relaxed edge in low-liquidity windows
Loss-burst cooldown	Per-asset 2-hour lockout after any 15M loss; prevents loss-clustering on the same asset
STC sizing scaler	Universal — reduces position proportionally to time remaining (contracts × 300/STC when STC > 300s); less time exposure = less risk
Price re-validation	Before every execution step, re-checks market conditions to confirm trade still makes sense

---

Edge Sources

The bot's expected profitability comes from four structural advantages:

Speed and Breadth of Information

While a typical Kalshi trader might check the Bitcoin price on one website, this system simultaneously processes real-time data from multiple professional-grade feeds. It detects cross-exchange price movements — where a large buy on one exchange precedes a move on another — and incorporates these signals before the Kalshi market adjusts.

Volatility Sophistication

The probability of a crypto asset staying above a given price depends critically on how much the price is expected to move. The bot uses academic-grade volatility estimation techniques (Realized Kernel estimators from Barndorff-Nielsen 2008, data-adaptive bandwidth selection, EGARCH conditional volatility, Mincer-Zarnowitz R^2-weighted blending) that are standard in institutional finance but rare among retail prediction market participants. This produces more accurate probability estimates, especially during volatile periods.

Distribution Fitting

Most quantitative models assume returns follow a simple bell curve (Gaussian) or a generic fat-tailed distribution. This system fits **Normal Inverse Gaussian distributions** to each cryptocurrency individually, capturing the specific tail behavior and asymmetry of BTC, ETH, SOL, and XRP. Statistical tests confirm NIG fits the actual data far better than generic alternatives.

Fee Optimization

Kalshi charges no fee on maker (limit) orders — taker fees apply only when the bot escalates to IOC execution. The bot's maker-first execution strategy avoids taker fees whenever possible, directly improving the profit margin on every trade. When forced to pay taker fees (locked spreads, SOL taker-first, decided contract overlay), the system re-verifies profitability before proceeding. This fee advantage compounds significantly over hundreds of trades.

Risk Management

Capital preservation is a core component of profit maximization. The system manages risk through multiple independent layers.

Conservative Position Sizing

The bot uses an **edge-tiered** approach — higher-conviction trades (greater model edge over the market) receive larger allocations, while lower-edge trades get minimal sizing. Individual trades risk a precisely calculated fraction of the bankroll. Even a string of losses has a limited impact on total capital. The sizing tiers were calibrated against actual trade performance data.

Automatic De-Risking

If the account balance drops below certain thresholds relative to its starting value, position sizes are automatically reduced:

Balance vs. 7-day rolling cash HWM	Action
≥ 85%	Full sizing
75–85%	**Half sizing**
65–75%	**Quarter sizing**
< 65%	**Trading halts**

This creates a geometric de-risking curve: the more the account loses, the less it risks, making recovery from drawdowns more manageable.

Multiple Safety Checks

Before any trade is placed, the system verifies:

1. The model's probability estimate passes a sanity check against the market price
2. The estimated edge exceeds the price-dependent minimum (0.25%–1.0%) after all fees (worst-case taker rates)
3. The contract price falls within acceptable bounds (80–99¢, with per-asset floors)
4. No extreme statistical indicators suggest unreliable model inputs
5. The position size respects all hard limits and drawdown adjustments

If any single check fails, the trade is refused — no exceptions. The system is designed to say "no" far more often than "yes."

Hard Price Boundaries

The bot only trades contracts priced between **75 and 99 cents**, with per-asset minimums (BTC 88¢, ETH 90¢ main tier with a separate 75–79¢ live sub-tier capped at 50 contracts, SOL 86¢, XRP 92¢). Below these floors, the probability of payout after fees is insufficient. Above 99¢, the potential profit is too small to justify the risk. Specialized overlays extend into lower prices near expiry: Terminal Momentum (96/98/99¢, final 1–5 min) and Low-Price Near-Expiry (BTC 80–87¢, final 10–120 sec). A SOL sub-86¢ time gate blocks low-price far-from-expiry entries. These guardrails eliminate an entire class of low-quality trades.

Intelligent Late-Window Execution

Below 180 seconds before settlement, the system switches to **direct taker execution** — skipping the maker order entirely and submitting an immediate-or-cancel order. Data showed maker fill rates of only 7.7% at low STC, meaning 92% of promising opportunities were missed. Direct taker captures these while still requiring full edge and liquidity validation.

Performance

Live Trading Results

Metric	Value
Status	Live trading since February 22, 2026
Settled trades	3,310 (3,066W / 242L / 2 BE)
Win rate	92.6%
Assets	BTC, ETH, SOL, XRP

Metric	Value
Entry prices	75–99¢ (per-asset: BTC 88¢+, ETH 90¢+ main tier with 75–79¢ capped sub-tier, SOL 86¢+, XRP 92¢+; overlays extend lower in the final minutes)

Market Expansion Pipeline

Beyond the live 15M engine, the platform has one already-live secondary side (weather NO in 1-contract verification mode), two active observation engines (S&P 500 intraday, sports comebacks), and one engine that was promoted, kill-switched, and now sits dormant pending re-enable (hourly crypto). Each followed (or is following) the same shadow → observation → live promotion pipeline.

Crypto Hourly Markets (Disabled)

Hourly cryptocurrency markets (KXBTCD, KXETHD, KXSOLD, KXXRPD) with 75 strikes per event. **Currently disabled (kill-switched Apr 18)**. The hourly system was briefly promoted to live trading in late February and reverted after analysis showed correlated multi-strike exposure was producing concentrated losses. Both `HOURLY_LIVE_ENABLED` and `HOURLY_NO_SIDE_LIVE` env vars must be flipped to 1 on the VPS to re-enable. The dedicated hourly CalEngine has been disabled due to +44pp overconfidence; a temperature scaling approach (T=1.45) is used instead when re-enabled. BTC was the only historically viable hourly asset; ETH/SOL are structurally unprofitable after fees at hourly timescales and XRP is fundamentally broken. Per-window position limits (max 2) and risk caps (15%) prevent correlated multi-strike losses on re-enable.

S&P 500 Intraday Markets

15-minute binary contracts on the S&P 500 index during NYSE regular trading hours (9:30 AM–4:00 PM ET). The SPX engine uses the same EGARCH volatility framework as crypto, adapted for equity-specific dynamics:

Adaptation	Detail
Leverage effect	Down moves increase equity volatility ~4× more than crypto — different EGARCH bounds
VIX integration	Forward-looking volatility signal blended with realized when they diverge significantly
Intraday seasonality	U-shaped pattern deseasonalized via 13 half-hour buckets to prevent time-of-day bias
Lower fees	Finance category — half the crypto fee multiplier (0.035 vs 0.07 taker), no maker fee
Correlation controls	Max 2 positions and 15% risk per window to prevent multi-strike blowups

This vertical leverages the same infrastructure while accessing a much larger and more liquid underlying market. It was briefly promoted to live trading on March 17 but reverted the same day after the primary price data feed (Polygon.io) began returning errors. Currently in observation mode with a fallback price feed, collecting calibration data with conservative eighth-Kelly sizing.

Weather Temperature Markets

Daily high temperature markets across **19 major US cities** — from New York and Chicago to Phoenix, Seattle, and New Orleans. This vertical is fundamentally different from financial markets — it uses **weather forecast ensembles** rather than price-based models:

Feature	Detail
82 independent forecasts	31 from NOAA's GFS and 51 from ECMWF (European weather model)
Probabilistic framework	Ensemble spread maps directly to outcome probability
Bias correction	Per-city learning system tracks and corrects forecast errors over time

Feature	Detail
Model-heavy blend	80% model / 20% market — ensemble forecasts are the primary signal

Research to date indicates the YES-side does not currently have a tradeable edge — the per-city Gaussian fit materially overestimates YES probability vs. actual outcomes. However, the **NO-side has been LIVE since 2026-04-11** in 1-contract verification mode: buying NO contracts at 37–40¢ (floor raised from 36¢ to 37¢ on May 1) when STC ≥ 16 hours before settlement. This exploits weather markets where the YES outcome is overpriced — making NO at 37–40¢ the profitable side. The 1-contract sizing reflects the verification-mode goal: collect outcome data on the bot's own NO entries (rather than counterfactuals) before any size promotion. Per-city CalEngines continue learning on every settlement to correct the model's biases. The YES-side remains observation only.

Live Sports Outcomes

Game outcome markets across **28 leagues** including NBA, NHL, MLB, NFL, EPL, ATP/WTA Tennis, and major soccer leagues worldwide. The sports engine identifies a specific high-value pattern: **pregame favorites trailing in-game**.

When a team or player that was heavily favored before the game falls behind, the market often overreacts — pricing the favorite far below its historical comeback probability. The engine uses a Bayesian model calibrated on historical comeback data to identify when the market discount is excessive:

Feature	Detail
28 leagues monitored	Binary (US sports, UFC, tennis) and three-way (soccer with draw)
Tennis support	ATP and WTA — player codes derived from names, set-based scoring
Conservative model	Likelihood ratios compressed 80% toward neutral to prevent overconfidence
Data collection mode	Entry criteria relaxed to capture wide range of scenarios for calibration
One signal per game	Prevents correlated exposure from multiple entries in the same game

Early results are promising for basketball specifically — the comeback model shows statistically significant edge (Fisher p=0.035) for NBA games. Other sport groups are closer to breakeven or negative. The engine is collecting more data through sequential statistical testing (SPRT) before a promotion decision, likely focused on basketball first rather than all sports simultaneously.

Expansion Philosophy

Each new vertical follows the same disciplined pipeline:

```
Build domain-specific model
         ↓
Shadow mode — log all signals without executing
         ↓
Calibration — track accuracy against outcomes over weeks/months
         ↓
Validation — promote only when data confirms genuine edge
         ↓
Conservative sizing — new verticals start with lower risk limits
```

This approach ensures each vertical is validated on real market data before real capital is deployed.

Infrastructure and Reliability

Always-On Operation

The bot runs on a dedicated cloud server (DigitalOcean, Ubuntu 24.04, 2GB RAM) managed by systemd, the standard Linux process manager. If the bot crashes for any reason, systemd automatically restarts it within seconds. The system has been designed for unattended 24/7 operation.

Continuous Deployment

Code changes pushed to the main branch automatically deploy to the production server via GitHub Actions:

1. The CI pipeline SSHes into the server
2. Pulls the latest code

3. Runs a syntax check to catch errors before they reach production
4. Restarts the bot service

This pipeline ensures rapid iteration while maintaining a safety net against broken deployments.

Complete Audit Trail

Every decision the bot makes is logged across three complementary systems:

System	What It Captures
SQLite database	Positions, orders, fills, settlements, every market evaluation with filter stage, full order lifecycle (order_id, submission time, final outcome)
JSONL journals	Append-only logs for scans, opportunities, rejections, trades, settlements, and maker fill model training data. Rotated daily with 30-day retention
Supabase dashboard	Real-time web interface (30s sync interval) showing positions, P&L, volatility, orderbooks, execution health, calibration diagnostics, and all shadow system data

This comprehensive logging enables full after-the-fact analysis of any trade or decision.

Crash Recovery

Order identifiers are written to the database before API submission. If the bot crashes mid-order, it can reconcile its state on restart without placing duplicate orders or losing track of open positions.

AI-Powered Analysis

An integrated analyst system (powered by Claude API) automatically examines every losing trade, identifies root causes, detects recurring patterns, and sends high-confidence findings via Telegram. This provides continuous diagnostic feedback without requiring manual review of every trade.

Technical Appendix

For readers interested in the mathematical foundations, the full technical whitepaper provides detailed formulas and derivations. Brief summaries of the key models:

Volatility Model — Uses the Realized Kernel estimator (Barndorff-Nielsen 2008) with data-adaptive bandwidth selection to produce noise-robust volatility from high-frequency returns. Multiple estimators are blended using Mincer-Zarnowitz R^2-weighted EMA blending, replacing fixed weights with data-driven quality scores. Includes EGARCH(1,1) with Student-t innovations for conditional volatility (promoted to live trading), time-varying RK weights, adaptive jump detection (percentile-based thresholds per asset), and options-implied volatility integration from Deribit.

Probability Model — Computes win probability using the Normal Inverse Gaussian (NIG) distribution with per-asset fitted parameters (a, b, μ, δ), capturing both heavy tails and asymmetry. NIG dramatically outperforms Student-t on statistical fit tests. Calibration is data-driven via per-product CalEngines: the 15M engine currently runs in passthrough mode (raw probability has lower Brier than the BLR fit, so the BLR layer is bypassed), while per-city weather, per-sport-group, and SPX-D engines run their full Platt → Beta → BLR pipelines. A dynamic time-dependent probability cap applies during startup (93% at 10min+ → 99.5% at <1min) but is bypassed (99.9% ceiling) once learned calibration is active. Final probability blends 60/40 (60% model, 40% market) for 15M; weather and SPX use product-specific weights.

Position Sizing — Edge-tiered sizing with drawdown-based scaling. Eight tiers from 25% at 4%+ edge down to 2% at 0.25%+ edge, with automatic de-risking against the rolling 7-day cash HWM (half at 85%, quarter at 75%, halt at 65%). Per-asset max risk per trade: BTC 15%, ETH 20%, SOL 15%, XRP 15%, hourly 15%, SPX 10%, weather 10%. 15M main uses full Kelly; hourly (when re-enabled) sizes at fixed 25 contracts on YES and 25 contracts on its DC overlay (bypassing Kelly); SPX uses eighth-Kelly (0.125); weather YES sim uses quarter-Kelly (0.25), weather NO is fixed 1-contract. Decided contracts use fixed sizing (T1/T1B/T2 at 20%, T2-Z25 at 10%, with SOL DC overrides at 5%/10% for ≥97¢/95–96¢). Low-STC sizing cap halves position below 100s; universal STC sizing scaler reduces position proportionally to time remaining (contracts × 300/STC) above 300s. LPNE is 50 contracts fixed.

Execution Model — Maker-first by default with three-tier post_only rejection handler: normal maker → degraded maker (1¢ worse) → taker IOC (with edge re-verification). Direct taker below 180s STC (data: 7.7% maker fill rate at low STC). Maker orders use `post_only=True` for no maker fee. SOL bypasses maker entirely (direct taker at all STC). BTC uses 7s escalation wait (vs 15s default). Decided contract overlay routes near-certain outcomes to direct taker — four live tiers (T1, T1B, T2 at 20% fixed; T2-Z25 at 10% after Apr 21 underperformance cut); T2-Z2 returned to shadow Apr 22; 6 expansion shadows collect data for future tiers. Terminal Momentum trades 96/98/99¢ in the final 1–5 min. Low-Price Near-Expiry intercepts BTC 80–87¢ at 50 contracts in the final 10–120 sec. Weekend (90¢+) and overnight (89¢+) discount overlays operate in low-liquidity windows with STC≤600s. Loss-burst cooldown locks each asset for 2 hours after any 15M loss. Fill detection via Kalshi WebSocket (zero API cost). Unfilled orders escalate via in-place amendment (`amend_order()`) before falling back to cancel + IOC (`time_in_force="immediate_or_cancel"`). Queue position monitoring every ~5s enables optimal escalation timing. Full order lifecycle tracking (order_id, submission time, outcome).

SPX Engine — Adapts the crypto EGARCH framework for S&P 500 equities: stronger leverage effect bounds (4× crypto), VIX-implied volatility integration when realized and implied diverge >30%, intraday seasonal deseasonalization (13 half-hour buckets), NYSE market hours guard with holiday calendar, half-rate fees (finance category), and per-window correlation controls (max 2 positions, 15% risk).

Weather Engine — Gaussian probability model over 82-member NWP ensemble (31 GFS + 51 ECMWF) across 19 US cities. Per-city EWMA bias correction with 7-day half-life. 80/20 model/market blend (ensemble is primary signal). Supports bracket, threshold, and tail probability market types.

Sports Engine — Bayesian comeback model using empirically calibrated likelihood ratios keyed on (deficit bucket, time remaining, pregame strength). Conservative LR scaling (80% compression toward neutral). 28 leagues with binary and three-way (soccer draw) outcome types, including ATP/WTA tennis with specialized player code derivation, set-based scoring, and gender-filtered ESPN parsing.

Last updated: 2026-05-02T12:18:34Z