BLUEHORNS AI

CRYPTO-NATIVE PORTFOLIO RISK INFRASTRUCTURE

Why Gaussian Value-at-Risk is statistically inadequate for digital-asset portfolios, and an empirically validated alternative

---

SCIENTIFIC WHITE PAPER

Market thesis, methodological foundations and independent out-of-sample validation

Confidential & Proprietary · Jean-Luc Caut · September 2026

Table of Contents

Executive Summary

Institutional capital continues to move toward digital assets, but the risk layer beneath that capital has, until now, been imported wholesale from traditional finance. This paper makes a narrower and more falsifiable claim than a generic “AI platform” pitch: the Gaussian-parametric assumptions embedded in classical Value-at-Risk (VaR) and Conditional VaR / Expected Shortfall (CVaR/ES), the metrics that anchor the current generation of institutional crypto-risk platforms, are measurably incompatible with the statistical properties of crypto-asset returns, and that incompatibility is documented in the peer-reviewed literature and independently reproducible on realized market data.

Two bodies of evidence support this claim. First, a body of academic work spanning 2018–2026 (Section 3.2–3.3) shows, across Bitcoin, Ethereum, Litecoin and Ripple and across more than one market cycle, that crypto returns carry excess kurtosis, negative skewness, volatility clustering and long-memory dependence that formal statistical tests Jarque-Bera, Shapiro-Wilk, and VaR-backtesting procedures such as Kupiec's and Christoffersen's reject as inconsistent with the normal distribution that parametric VaR/CVaR assume.

Second, an independent replication performed for this paper (Section 3.4) on 512 trading days of realized BTC-USD outcomes shows that a textbook Gaussian, square-root-of-time VaR band, built from the same realized-volatility input a parametric risk system would use, breached its stated 95% confidence level 56% more often than it should have, and breached disproportionately from low-volatility starting conditions: precisely when a parametric model is most likely to understate emerging risk.

Sections 4–5 present BlueHorn's methodological response an ensemble, tail-aware, dual-head TCN-Transformer architecture calibrated with split conformal prediction rather than a Gaussian quantile function and report its performance on a fully out-of-sample, hash-verified 17-month BTC-USD test window the model never saw during training or calibration.

That forecasting engine is one of four integrated capabilities, behavioural risk clustering, multi-dimensional asset rating, probabilistic forecasting and simulation-based portfolio optimization, underpinned by a cryptographically verifiable trust layer, described in full in Section 6.

The paper reports these results in full, including the findings that qualify the headline numbers: the model's raw point forecast does not beat a naive no-change baseline, and its high-conviction trading signal is short-biased in this particular window.

This is a methodological choice, not an oversight. A platform whose only public evidence is a single favourable accuracy figure is indistinguishable, under due diligence, from a platform with no evidence at all; the figures that matter to a CRO or a model-risk committee are calibration, drawdown and risk-adjusted return under a fully specified cost model, not a headline MAPE.

The paper also revisits the addressable market (Section 2.1–2.2) and the competitive landscape (Section 2.3) with the same discipline. The crypto portfolio-risk category already has credible incumbents: Talos following its acquisition of Cloudwall, Kaiko and Amberdata, and the addressable software TAM is presented as a transparent, sourced scenario range rather than a single unexplained figure.

BlueHorn's differentiation is architectural and empirical: a risk engine built for, and validated against, the actual statistical behaviour of crypto-asset returns, rather than one that imports a risk vocabulary built for Gaussian, low-frequency, exchange-cleared equity and rates markets.

1. Purpose and Scope

This paper addresses risk-management infrastructure for portfolios composed of crypto-assets specifically spot and derivative exposure to native tokens, stablecoins and on-chain instruments rather than the broader institutional “digital assets” category that also spans tokenized securities, digital cash and CBDC-adjacent instruments.

Where the paper cites general institutional-AUM or digital-asset-allocation statistics (Section 2.1), those figures are used only to size the addressable capital base from which crypto-specific risk-software demand can emerge; they are not presented as a measurement of crypto-portfolio AUM itself, which is not separately or reliably reported in any public dataset known to the authors.

The paper's central technical claim is deliberately narrow and falsifiable: that the Gaussian-parametric implementations of VaR and CVaR embedded in the current generation of commercial crypto-risk tooling are statistically incorrectly calibrated for crypto-asset return distributions.

It is not claimed that Value-at-Risk as a concept is invalid, nor that every implementation offered by every incumbent is Gaussian — several, per their own public materials, use historical-simulation or exponentially-weighted variants. Section 3.5 addresses each implementation family on its own terms.

Structurally, the paper proceeds from market context (Section 2), to the statistical case against Gaussian risk metrics for crypto portfolios (Section 3), to BlueHorn's methodology (Section 4)

and its independent, out-of-sample empirical validation (Section 5), to product, go-to-market and commercial architecture (Sections 6–7).

2. The Crypto-Native Portfolio-Risk Software Market

2.1 Institutional capital base and the crypto allocation trajectory

Global asset-management AUM reached approximately $147 trillion in 2025 (McKinsey; BCG). This is the broad institutional capital base from which crypto allocations can emerge — it is not itself a crypto figure. The OECD separately reports $69.8 trillion of pension assets across OECD and non-OECD jurisdictions at end-2024, and the Thinking Ahead Institute reports $68.3 trillion of global pension assets in 2025; these two pension series overlap with each other and with the $147tn figure and must not be summed.

CoinShares' August 2026 Digital Asset Fund Manager Survey covered 30 institutional respondents representing approximately $1.16 trillion of AUM and reported an average digital-asset allocation of 1.2%, implying roughly $13.9 billion of digital-asset exposure within that respondent pool. It is a useful, current empirical anchor, but it is a survey of self-selected respondents, not a census, and it is not restricted to portfolios composed exclusively of crypto-assets.

State Street's 2025 research reports a broader current digital-asset allocation of approximately 7%, with a three-year forward target of approximately 16% among surveyed institutions — but State Street's definition of “digital assets” explicitly includes tokenized securities and digital cash alongside crypto-assets, so this series is directionally informative about institutional intent, not a pure-crypto allocation estimate.

A Coalition Greenwich study found that 53% of surveyed managers cited portfolio-management and risk applications specifically as a digital-asset infrastructure need, and CoinShares' 2025–2026 survey series has repeatedly identified custody, accessibility and corporate/regulatory restriction, not analytical capability alone, as the leading constraints on further allocation. Read together, these findings support treating crypto-native risk infrastructure as part of the institutional adoption bottleneck, not a reporting convenience layered on top of an already-solved allocation problem.

2.2 A disciplined software-TAM model

A defensible way to size the addressable software opportunity, without conflating it with total institutional AUM, is to apply a specialist risk-platform fee to a range of plausible crypto-

allocation scenarios against the $147tn AUM base. Table 1 presents three allocation scenarios (0.5%, 1.0%, 2.0%) and two illustrative fee bands (5 and 15 basis points, broadly consistent with specialist institutional risk-platform pricing) as an analytical sensitivity, not a market-research estimate of realized revenue.

Scenario (crypto allocation)	Implied crypto AUM	Software TAM @ 5 bps	Software TAM @ 15 bps
0.5% of global AM AUM	$0.73tn	$0.37bn	$1.10bn
1.0% of global AM AUM	$1.47tn	$0.74bn	$2.21bn
2.0% of global AM AUM	$2.94tn	$1.47bn	$4.41bn

Table 1. *Illustrative software-TAM sensitivity. Crypto-AUM scenarios apply 0.5–2.0% allocation rates to $147tn of global asset-management AUM (McKinsey, 2025; BCG, 2026); the scenario is intentionally broader than pure-crypto funds and should be read as an upper-bound addressable pool, not observed pure-crypto AUM. Fee bands are illustrative and will depend in practice on minimum platform fees, portfolio count, API volume and deployment model.*

2.3 Competitive landscape: a fragmented, not empty, category

The defensible competitive position is not that the category is empty.

Talos acquired **Cloudwall** in April 2024 specifically to add digital-asset portfolio-risk technology, real-time P&L, VaR/CVaR, stress testing and portfolio construction to its execution and portfolio-management stack.

Kaiko markets an institutional portfolio-risk and performance-analytics suite built on historical-simulation VaR with volatility-decay weighting.

Amberdata offers portfolio and risk management with daily VaR and classical parametric analytics designed to let institutional clients apply familiar traditional-finance frameworks to digital assets.

The defensible claim is that the category remains fragmented and methodologically undifferentiated: no identified incumbent combines, in one crypto-native decision layer, behavioural risk-state detection, continuous multi-dimensional asset ratings, calibrated probabilistic forecasting, tail-aware scenario optimization and a cryptographically verifiable decision-audit trail.

Sections 6.1–6.5 set out each of these five elements in turn, together with the institutional evidence standard a buyer should require of each (Table 9).

Provider	Core strength	Risk-engine basis	Key gap vs. BlueHorns thesis
Talos / Cloudwall	Institutional PMS/OEMS + risk	VaR/CVaR, scenario weighting	Risk is embedded in broader lifecycle infrastructure; no behavioural risk clustering or continuous asset ratings of its own (Sec. 6.1–6.2)
Kaiko	Institutional market data + portfolio risk	Historical-simulation VaR, decay-weighted	Strong data position; historical-simulation VaR remains exposed to regime non-stationarity (Sec. 3), and there is no simulation-based optimization layer (Sec. 6.4)
Amberdata	On/off-chain data + risk analytics	Daily VaR, classical parametric analytics	Broad data intelligence; parametric VaR inherits the normality violation directly (Sec. 3), with no continuous asset-rating layer (Sec. 6.2)
Glassnode	On-chain / market intelligence	Not a portfolio-optimization layer	Excellent signal layer; not a complete institutional decision-audit system (Sec. 6.5)
MSCI / Barra; BlackRock Aladdin	Traditional factor/portfolio risk at scale	Gaussian factor models built for equities/rates	Deep distribution; Gaussian factor models built for equities/rates require ground-up adaptation for 24/7 markets and crypto microstructure, including a behavioural clustering layer they do not offer (Sec. 6.1)

Table 2. *Competitive positioning. Descriptions of competitor functionality reflect each vendor's own public materials and are treated as company claims, not independently verified performance assessments.*

3. Why Classical VaR and CVaR Are Statistically Inadequate for Crypto

3.1 The distributional assumptions embedded in classical risk metrics

The dominant institutional risk metric, Value-at-Risk, is a quantile of the portfolio loss distribution: VaR_p is the loss level such that $P(Loss > VaR_p) = p$ over a stated horizon. In its parametric ("Delta-Normal" / variance-covariance) implementation, the form embedded in RiskMetrics-style engines and in the classical parametric analytics several incumbent platforms describe in their own materials, the quantile is computed by assuming returns are normally distributed, $r \rho N(\mu, \sigma^2)$, so that $VaR_p = \mu + \sigma \cdot \Phi^{-1}(p)$, and multi-day horizons are obtained by scaling single-day volatility with the square-root-of-time rule, $\sigma_n = \sigma_1 \cdot \sqrt{n}$. Both steps require i.i.d. Gaussian (or at minimum, thin-tailed and temporally homogeneous) return innovations. Conditional Value-at-Risk / Expected Shortfall (CVaR/ES) is the expected loss conditional on breaching VaR_p; it improves on VaR by describing the tail rather than only its boundary, but a parametric CVaR inherits the same distributional assumption, and even a historical-simulation

CVaR assumes the historical estimation window is representative of the near-term future — an assumption that volatility-clustering and regime-shift dynamics (Section 3.2) directly violate.

None of this is a crypto-specific critique in principle: the fragility of Gaussian VaR under fat tails is a known issue in traditional-finance risk management, which is precisely why extreme-value-theory (EVT), GARCH-family and historical-simulation alternatives to naive Delta-Normal VaR already exist. The claim advanced here is empirical and asset-class-specific: crypto-asset returns depart from normality, and depart from the temporal homogeneity the square-root-of-time rule requires, by a margin large enough and consistent enough across assets, sample periods and studies that Gaussian-parametric VaR/CVaR should be treated as unreliable for crypto portfolios by default, not merely as an approximation with a known error bar.

3.2 Published evidence: crypto returns are fat-tailed and long-memory

A 2025 study of Bitcoin, Ethereum, Litecoin and Ripple daily log-returns from August 2015 to July 2022 reports excess kurtosis of 9.78 (Bitcoin), 18.58 (Ethereum), 11.44 (Litecoin) and 24.99 (Ripple) against a Gaussian benchmark of zero, with negative skewness for Bitcoin and Ethereum and positive skewness for Litecoin and Ripple, asymmetric tail risk that a symmetric Gaussian band cannot represent in either direction. Jarque-Bera and Shapiro-Wilk normality tests reject the null of normality for all four series at $p < 0.0001$. The same study documents long-memory (Hurst exponent above 0.6, rising above 0.65–0.72 for squared returns) and significant ARCH effects across all four assets, meaning volatility shocks decay slowly and cluster in time rather than reverting quickly, the structural reason the square-root-of-time scaling rule is unreliable for crypto (Subramoney, Chinhamu & Chifurira, 2025).

This is consistent with, and extends, an earlier line of evidence. Applying extreme-value theory to five major cryptocurrencies, Gkillas and Katsiampa (2018) found tail behaviour far heavier than a normal or even a standard fat-tailed parametric distribution would predict, with substantial variation in tail risk across assets. Zhang, Li, Xiong and Wang (2021) document a systematic relationship between downside risk, measured through VaR, expected shortfall and downside beta, and the cross-section of subsequent crypto returns, implying that tail behaviour is not merely a modelling nuisance but a priced, economically meaningful feature of these markets. Kurosaki et al. (2021) show that multivariate normal tempered stable processes with GARCH dynamics fit crypto return data materially better than Gaussian alternatives. Each of these studies uses different assets, sample windows and estimation methods, and they agree on the direction of the finding.

3.3 Published evidence: VaR/CVaR backtests fail on crypto assets

The more decisive evidence is not distributional, it is what happens when a VaR model is actually backtested against realized outcomes. Backtesting compares each period's stated VaR quantile against the realized loss and counts “violations”: exceedances of the stated threshold. Under a correctly calibrated 95% VaR model, violations should occur in almost exactly 5% of periods, independently distributed through time, formally verifiable with the Kupiec (1995) unconditional-coverage test and the Christoffersen (1998) conditional-coverage/independence test.

Applying both tests in-sample and out-of-sample to Bitcoin, Ethereum, Litecoin and Ripple, Subramoney, Chinhamu and Chifurira (2025) find that standard GARCH models with Gaussian innovations are stable but “tend to under-represent extreme risk events,” and that only long-memory, heavy-tailed extensions (FIAPARCH combined with generalized hyperbolic or generalized lambda innovation distributions) pass both back tests reliably at the extreme quantiles institutional risk management cares about most. A parallel literature reaches the same conclusion by a different route: hybrid GARCH-EVT-copula models are repeatedly reported to outperform historical-simulation and variance-covariance VaR/CVaR for Bitcoin, Ripple and other crypto assets under formal backtesting, and Alexander and Dakos (2023) show that even exponentially-weighted-moving-average VaR/ES models, already an improvement over static Delta-Normal, require crypto-specific recalibration to achieve acceptable accuracy for crypto portfolios.

Taken together, this is not a single contested result: it is a consistent finding, replicated across at least four independent research groups, three risk-metric families (VaR, CVaR/ES, range-VaR) and two decades' worth of overlapping but non-identical sample windows, that Gaussian and naive-historical implementations of classical risk metrics fail formal back tests on crypto-asset returns, and that the models which do pass are specifically the ones built to accommodate fat tails, asymmetry and long-memory volatility, features standard commercial VaR/CVaR engines do not, by design, model.

3.4 Independent replication on 512 days of realized BTC-USD outcomes

To ground Sections 3.2–3.3 in data specific to this paper rather than relying solely on third-party samples, we replicated the core logic of a Gaussian, square-root-of-time VaR band directly on the 512-trading-day, hash-verified out-of-sample BTC-USD window used for BlueHorn's own model evaluation (Section 5.1), using the realized daily volatility already present in that release as the sole volatility input, the same input a parametric risk engine would use.

This isolates the distributional-assumption question from any question about BlueHorn's own model quality: the band tested below is a textbook Gaussian VaR construction, not a BlueHorns output.

Replication design

For each of the 512 trading days (7 April 2025 – 31 August 2026), a symmetric 95% band was constructed as $\pm 1.96 \cdot \sigma_1 \cdot \sqrt{3}$ around a zero-drift random walk assumption — the standard square-root-of-3-day scaling of the realized daily volatility — and compared against the realized 3-day-ahead log-return. "Breach" means the realized return fell outside the band. The identical realized-return series was used to evaluate the empirical coverage of BlueHorn's own split-conformal 95% interval (Section 5.2), enabling a like-for-like comparison on one dataset.

Metric	Naive Gaussian 95% VaR band	BlueHorn conformal 95% interval
Empirical coverage (target: 95%)	92.19%	98.44%
Breaches (of 512 trading days)	40 (7.81%)	8 (1.56%)
Breach rate vs. 5% nominal budget	1.56× (under-covers)	0.31× (over-covers)
Kupiec unconditional-coverage LR statistic	7.33	17.22
Kupiec test p-value (H_0: correctly calibrated)	0.0068 — reject	< 0.0001 — reject
Direction of miscalibration	Dangerous (breaches too often)	Conservative (breaches too rarely)

Table 3. *Naive Gaussian VaR vs. BlueHorn's conformal interval, identical 512-day out-of-sample BTC-USD window. Both bands formally fail an exact Kupiec test, but in opposite directions: the Gaussian band is unsafe (it under-states risk, which is the direction that produces surprise losses against a stated limit), while the conformal band is conservative (it over-states risk, which costs capital efficiency but not risk-limit integrity). Kupiec's test methodology follows Kupiec (1995).*

The regime breakdown is, if anything, more informative than the headline breach rate. Splitting the sample into low-, mid- and high-realized-volatility terciles shows the Gaussian band's breach rate falling as volatility rises, the opposite of what a naive "tail risk = high-volatility periods" intuition would predict.

Realized-volatility regime	Days	Gaussian band breach rate	Conformal interval breach rate
Low-volatility tercile	171	14.04%	1.75%
Mid-volatility tercile	170	7.06%	1.76%
High-volatility tercile	171	2.34%	1.17%

Table 4. *Breach rate by realized-volatility tercile, same window and construction as Table 3.*

This pattern is exactly what volatility clustering and regime persistence predict, and is consistent with the long-memory findings surveyed in Section 3.2: a trailing realized-volatility input systematically understates near-term risk precisely when a low-volatility regime is about to break, and overstates it once a high-volatility regime is already underway and beginning to mean-revert. A model calibrated on the assumption of i.i.d. Gaussian innovations has no structural mechanism to anticipate this asymmetry; a model calibrated on realized, regime-conditional non-conformity, as in split conformal prediction, does not require that assumption to begin with.

Finally, the realized 3-day log-return series itself, tested directly, rejects normality by a wide margin: skewness −0.31, excess kurtosis 4.08 (Gaussian = 0), a Jarque-Bera statistic of 363.2 ($p = 1.3\times10^{-79}$) and a Shapiro-Wilk statistic of 0.960 ($p = 1.3\times10^{-10}$). These figures are computed on a different asset window (BTC-USD, 3-day overlapping returns, April 2025–August 2026) and a different horizon than the daily-return statistics surveyed in Section 3.2, and they point in the same direction, negative skew, materially positive excess kurtosis, overwhelming rejection of normality, which is itself a form of replication across time and horizon, not a restatement of the same sample.

3.5 Implications for the incumbent vendor stack

This evidence does not support a blanket claim that “VaR/CVaR are invalid for crypto”; it supports a precise claim about which implementations are exposed and why. Parametric / variance-covariance VaR and CVaR, the form several incumbents describe using for at least part of their crypto-risk stack, inherit the normality violation in Section 3.2 directly: the quantile function itself assumes a distribution the data reject. Historical-simulation VaR, which does not assume normality, instead assumes the historical estimation window is stationary and representative of the near-term future; Section 3.4's regime-tercile results show precisely why that assumption is fragile for crypto, since the largest miscalibration occurs at regime transitions rather than within stable regimes. Exponentially-weighted variants improve on static historical simulation by unweighting recent data, but still require crypto-specific recalibration to pass formal back tests, per Alexander and Dakos (2023).

The practical consequence for institutional buyers is that the choice of risk-engine family is not a stylistic preference: it is a testable, falsifiable property of the platform, and it should be proceed the same way, with a Kupiec/Christoffersen backtest on the buyer's own portfolio and horizon, as any other quantitative claim a vendor makes. Section 5 applies exactly this standard to BlueHorn's own forecasting engine.

4. BlueHorn's Methodology: Tail-Aware, Calibrated Forecasting

BlueHorn's response to Section 3 is architectural rather than cosmetic: replace the Gaussian quantile function at the centre of classical VaR/CVaR with (i) a forecasting model whose error distribution is never assumed to be Gaussian, and (ii) a calibration layer, split conformal prediction, whose finite-sample coverage guarantee does not depend on a normality assumption at all.

4.1 Architecture: five-run ensemble, dual-branch, dual-head

The evaluated system ("TCN_Transformer_Dual_Head") targets the 3-day-ahead log-return of BTC-USD from a 7-day input window. Two branches run in parallel on the same input: a temporal-convolutional (TCN) branch, two residual, causally-padded, dilated 1-D convolution blocks with gated-linear-unit-style activation, layer normalization and spatial dropout, and a causal Transformer branch, a learned positional embedding followed by one causally-masked multi-head self-attention block (4 heads) and a GELU feed-forward sub-layer. The two branch outputs are combined through a learned, input-dependent fusion gate rather than a fixed weighting, then passed through pooling and dense layers to a shared representation.

That representation feeds two heads trained jointly: a 3-quantile head (q0.025, q0.50, q0.975) trained with a pinball loss plus a directional penalty on the median, and a binary head trained with cross-entropy to estimate the probability the 3-day return is positive. Five independently seeded instances are trained (AdamW, early stopping on validation quantile loss) and combined into an inverse-validation-loss-weighted ensemble, which is materially more robust to single-seed variance than any individual run.

4.2 Feature panel and leakage controls

The model consumes 31 daily features in three groups, 11 technical (moving-average deviations, RSI, realized volatility, MACD, Bollinger position/width, volume and returns), 10 on-chain (exchange flows, whale ratio, liquidation volumes, MVRV, Puell multiple, taker buy/sell ratio, Coinbase premium) and 10 macro (DXY, equity and gold returns, the US yield curve, VIX), joined with forward-fill only, so no feature can see information from after the date it is dated.

The chronological 70/15/15 train/validation/test split inserts a 3-day embargo at each boundary specifically to remove overlap between adjacent 3-day-forward targets, a standard safeguard against leakage in overlapping-horizon time series that is frequently absent from published crypto-forecasting studies.

4.3 Conformal calibration instead of a Gaussian quantile function

After ensembling, the combined quantile interval is calibrated post hoc with split conformal prediction (Vovk, Gammerman & Shafer, 2005) at the 95% level: a single additive correction, equal to the (1−α)-quantile of the validation-set non-conformity score, is added to the upper bound and subtracted from the lower bound of the test-set interval.

Unlike a Gaussian VaR band, this correction makes no distributional assumption about the return-generating process; its finite-sample coverage guarantee instead requires exchangeability between the calibration and test residuals, an assumption that is itself not strictly satisfied by regime-switching financial time series, and which Section 5.6 addresses explicitly rather than treating as automatically satisfied.

5. Empirical Validation: Out-of-Sample Case Study on BTC-USD

5.1 Experimental protocol and reproducibility

All figures in this section are computed from a released, SHA-256-checksummed test-set prediction file (512 rows) cross-checked against the ensemble's release manifest, which independently records a hash of every hyper parameter so that a downstream backtest cannot silently run on a stale or mismatched file.

The test window runs from 7 April 2025 to 31 August 2026, approximately 1.40 years, 512 trading days, and was never used for training or for fitting the conformal correction, which was calibrated on the validation split only. Section 3.4's independent replication was computed directly from this same released file, providing an internal cross-check between the two analyses in this paper.

5.2 Probabilistic calibration results

Figure 1 plots realized BTC-USD close against the ensemble's calibrated median forecast and 95% conformal interval across the full test window, including the sharp January–February 2026 drawdown and the late-August 2026 spike, both of which are visible as periods of interval widening.

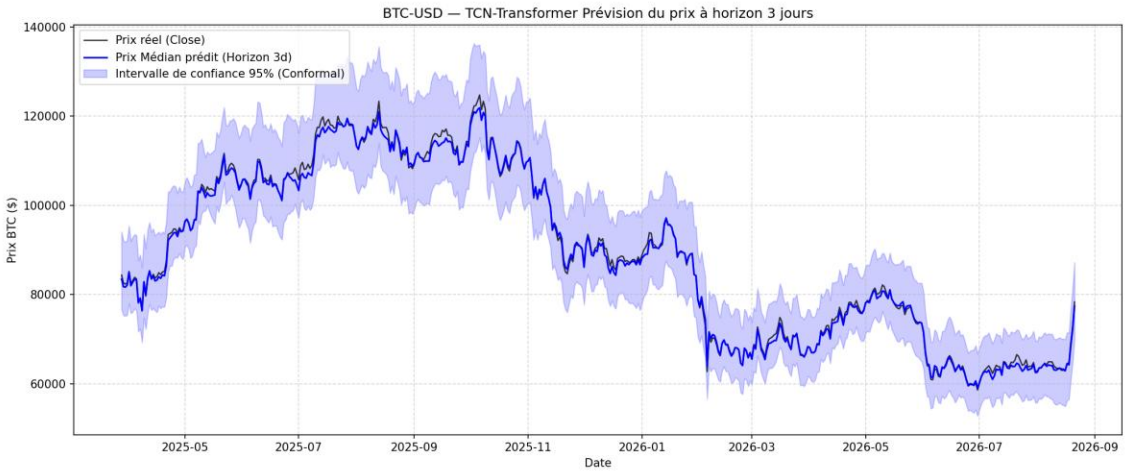

Figure 1. *Realized BTC-USD close vs. calibrated median forecast and 95% conformal interval, full 512-day out-of-sample test window.*

Metric	Value
Mean pinball loss (3 quantiles)	0.00672
Pinball loss, q0.025	0.00275
Pinball loss, q0.50 (median)	0.01445
Pinball loss, q0.975	0.00296
Empirical coverage of 95% interval	98.4% (nominal 95%)
Mean interval width (log-return)	20.6% (range 11.3–25.0%)
Breach split (below / above interval)	0.78% / 0.78%

Table 5. *Probabilistic accuracy on the 512-day out-of-sample test set.*

The interval is over-covered rather than under-covered, 0.78% of realizations fall below the lower bound and 0.78% above the upper bound, versus 2.5% expected on each side under exact calibration.

Combined with a mean width of 20.6% in log-return space (roughly ±10% around the median price forecast), this indicates the calibrated interval is valid but not sharp: comfortably wide enough to contain the true 3-day move, at the cost of some discriminative usefulness as a stand-alone risk bound. We report this qualification rather than the coverage figure alone, consistent with the standard that Section 3.5 asks incumbent vendors to meet.

5.3 Point-forecast and directional accuracy — a full accounting

Metric	Value
Median-forecast MAE (log-return)	2.89%
Median-forecast RMSE (log-return)	3.93%
Median-forecast MAPE (price, H+3)	2.89%
Naive persistence baseline MAPE (H+3)	2.78%
Directional accuracy — consensus-gated signals only (23.2% of days)	62.2%
Monthly raw directional accuracy — mean (std. dev.)	48.7% (10.7 pp)
Monthly raw directional accuracy — range across 17 months	29.2% – 66.7%

Table 6. *Point-forecast and directional metrics, same 512-day test set.*

Two results qualify the headline probabilistic performance, and we state them without euphemism. First, the median 3-day-ahead price forecast is marginally less accurate, in pure MAPE terms, than a trivial no-change baseline (2.89% vs. 2.78%): at this horizon, BTC-USD behaves close to a random walk, consistent with weak-form market-efficiency arguments frequently invoked in this literature (Section 5.5).

Monthly raw directional accuracy averages 48.7%, statistically indistinguishable from chance, with a 10.7-percentage-point standard deviation and a 29.2–66.7% range across the 17 calendar months in the test window, underlining that any apparent edge in the raw median forecast is unstable across regimes rather than a stationary model property.

Second, the 62.2% directional-accuracy figure applies only to the 23.2% of days on which the ensemble's volatility-scaled dead zone and weighted vote reach a ≥40% consensus. On those higher-conviction days the model is materially more informative, but the signal is heavily asymmetric: 118 “Sell” calls against a single “Buy” call over 512 days, even though the realized label was “Up” on 269 days (52.5%) and “Down” on 243 (47.5%).

The near-balanced realized outcome indicates the signal's short bias is a property of this model/threshold combination on this specific run, not of the underlying market during the test window, and it should be read as a regime-specific artefact requiring re-validation, not as a general finding.

Why report a weak point-forecast result in an investor-facing white paper?

Because the alternative — reporting only the interval-calibration and backtest results while omitting Table 6 — would itself misstate what the model does well. A 3-day crypto point-forecast that is not

more accurate than “assume tomorrow looks like today” is an expected, literature-consistent result, not a disqualifying one; it is the calibrated uncertainty (Section 5.2) and the risk-managed capital allocation built on top of it (Section 5.4) that constitute the actual product, and reporting Table 6 in full is what allows a model-risk reviewer to verify that distinction rather than take it on faith.

5.4 From signal to portfolio outcome: the risk-managed backtest

The prediction file was replayed through a rule-based execution engine that converts the consensus-gated signal into positions with volatility-scaled sizing (up to 90% of notional), a 2.5% hard stop-loss, a 2.0% trailing stop from each trade's peak, a two-day cooldown after two consecutive hard stops, and a 0.07% one-way transaction cost (0.05% fee plus 0.02% slippage).

Metric	Strategy	Buy & hold
Cumulative return	+26.14%	−0.87%
CAGR	18.00%	−0.62%
Annualized volatility	12.75%	42.17%
Sharpe ratio	1.36	−0.01
Sortino ratio	1.91	−0.02
Calmar ratio	2.31	−0.01
Maximum drawdown	−7.80%	−53.06%
Win rate (active days)	49.4%	n/a
Profit factor	1.41	n/a
Active days (position ≠ 0)	178 / 512 (34.8%)	n/a

Table 7. *Backtest performance, net of transaction costs, same 512-day out-of-sample window (17 months, April 2025–August 2026).*

A simplified, sizing-free replication (±1 unit exposure on every non-neutral signal, no stops) produces a 61.3% net-of-fee win rate on the 119 non-neutral trades, closely matching the 62.2% raw directional accuracy of the filtered signal in Table 6, an internal consistency check that the strategy P&L is an arithmetically coherent transform of the directional signal rather than an artifact of the position-sizing engine.

The improvement from that simplified backtest (Sharpe 1.23, max drawdown −28.7%) to the full engine (Sharpe 1.36, max drawdown −7.8%) is attributable almost entirely to the trailing-stop and volatility-scaled-sizing overlay, not to additional predictive information, since both use identical underlying signals.

The correct reading of Table 7 is therefore not "the model predicts Bitcoin", Section 5.3 already shows it does not, on a raw basis, but that a calibrated uncertainty estimate, filtered to a minority of higher-conviction days and combined with mechanical risk controls, converts a directionally weak signal into a capital-preservation outcome: 87% less maximum drawdown than buy-and-hold, at roughly one-third the realized volatility, over a window that included a >45% peak-to-trough Bitcoin decline. This is the risk-management value proposition Section 1 sets out to test, evaluated on its own terms.

5.5 Positioning against the published forecasting literature

No two published studies in this space share the present asset window, cost assumptions or validation protocol, so Table 8 is a qualitative positioning exercise, not a controlled benchmark.

Study	Architecture	Headline metric(s)	Protocol note
This work	5-run ensemble TCN + causal Transformer, dual quantile/direction heads, split-conformal 95%	Filtered directional acc. 62.2% (23% of days); backtest Sharpe 1.36	Single chronological split
Yousefnezhad et al. (2026)	Multi-scale TCN, on-chain + sentiment features	AUC 0.632; profit factor 1.70 vs. 5 baselines	5-fold walk-forward
Hasanli & Dursun (2026)	Transformer-LSTM hybrid vs. LSTM/GRU/BiLSTM/BiGRU	Lowest MAE/RMSE/MAPE among tested models	Multi-step 7/14/21-day
Cui et al. (2026)	LSTM/GRU direction signals + LLM reasoning + sentiment	F1 78.9% (direction); +194% return vs. B&H	12-month window
11-architecture study (2026)	Autoencoder-LSTM best of 11 archs vs. ARIMA	RMSE −9–15% vs. LSTM/GRU/RNN; deep hybrids overfit	2014–2024, BTC

Table 8. *Indicative, non-controlled comparison with recent hybrid architectures for Bitcoin forecasting; full references in the References section.*

Two comparisons are worth drawing out. First, the AUC/profit-factor pairing Yousefnezhad et al. (2026) report for a similarly on-chain-fed TCN is broadly consistent in shape with the gap observed here between near-chance raw directional accuracy and a materially more useful filtered/backtested signal; both studies separate tradable from non-tradable predictions with a volatility- or profit-based threshold rather than trading every forecast.

Second, the caution raised by the eleven-architecture comparison, that stacking additional components can over fit relative to a simpler, well-tuned baseline, applies directly to the dual-branch, multi-task design evaluated here: this release does not compare the dual-head ensemble against a simpler single-branch or single-task variant on the same split, so the

incremental value of the added architectural complexity is not established by this evaluation alone.

5.6 Limitations and the validation roadmap

We state the following limitations as BlueHorn's own near-term validation roadmap, not as objections raised against the platform by a third party:

- **Point-forecast skill is weak at this horizon.** The raw 3-day median forecast does not beat a naive persistence baseline, and raw directional accuracy is at chance. Economic value comes entirely from the ensemble-vote/dead zone filter, not the point forecast, this should be independently verified by any prospective buyer, not taken on the strength of Table 5 alone.
- **The filtered signal is regime-specific and short-biased in this run.** 118 of 119 non-neutral signals were “Sell” despite a near-balanced realized up/down split; monthly accuracy swings between 29% and 67%. This should not be assumed to generalize to other market regimes without re-validation.
- **Conformal coverage is valid here, but the guarantee is a heuristic for this data-generating process.** Split conformal prediction's finite-sample coverage guarantee formally requires exchangeability of calibration and test residuals, an assumption not strictly satisfied by regime-switching financial series. The strong empirical coverage observed here (Table 5) is a single-window result, not a proof the guarantee will hold indefinitely.
- **Single-path backtest; no walk-forward re-validation yet.** Unlike the five-fold walk-forward protocol used by Yousefnezhad et al. (2026), this evaluation uses one chronological split. No Diebold-Mariano or Model Confidence Set test against a baseline forecaster has yet been run, and no classical or simpler neural baseline (ARIMA, single-branch LSTM) has been evaluated on this exact split for a controlled comparison.
- **Backtest P&L conflates forecasting skill with risk-engineering.** The improvement from the sizing-free backtest to the full engine (Section 5.4) is driven by the risk overlay, not by additional predictive information, since both use identical signals, a distinction every institutional reviewer of Table 7 should independently confirm.
- **Trading-cost assumptions are simplified.** A flat 0.07% one-way cost is applied; funding or borrow costs for sustained short exposure, partial fills and slippage beyond the fixed assumption are not modelled, which likely overstates achievable net performance at the position sizes used (up to 90% notional).
- **Single-asset, single-horizon scope.** This case study covers BTC-USD at a 3-day horizon only. Enterprise deployment requires the same protocol, hash-verified, leakage-controlled, walk-

forward, cost-inclusive, replicated across the full asset universe, horizon set and stress periods a given mandate requires.

This is precisely the validation package institutional model-risk functions should require of any vendor, rolling-origin back tests, untouched test periods, crisis subsamples, multiple baselines, confidence intervals and statistical tests, full data/feature lineage, and it is the standard against which Section 3.5 asks incumbent vendors' own VaR/CVaR engines to be diligenced in turn.

6. Product Architecture and Institutional Buyer Fit

BlueHorn's architecture is presented deliberately as an institutional quantitative risk system, not an “AI platform”, a framing intended to be legible to CIOs, CROs, portfolio managers and model-risk committees rather than to read as a generic AI wrapper.

The platform is organized around four capabilities that operate as a single pipeline, so that a pattern surfaced at one stage feeds directly into the next: behavioural risk clustering identifies hidden common exposures across a portfolio, multi-dimensional asset rating translates that structure into a common risk vocabulary, probabilistic trend forecasting supplies calibrated, tail-aware directional and uncertainty estimates, and simulation-based portfolio optimization turns all three into an executable allocation. A cryptographic trust layer underpins all four, time-stamping and verifying every decision the pipeline produces. Sections 6.1–6.5 describe each capability in turn; Table 9 summarizes the institutional evidence standard attached to each layer, and Section 6.7 sets out the primary buyer personas each one is built to serve.

6.1 Behavioural Risk Clustering

Rather than grouping assets by sector label, the platform groups them by how they actually behave, their volatility, drawdown and recovery characteristics, often revealing that assets from unrelated categories share the same underlying risk profile. This is used to flag what the team calls "false diversification": portfolios that appear varied on paper but are, in practice, concentrated in a single hidden risk archetype.

Because the clustering is unsupervised and refreshed continuously on incoming market and behavioural data, rather than fixed at onboarding, it is designed to surface regime-driven convergence, assets that behaved independently in one volatility regime and jointly in the next, the same non-stationarity that Section 3.4's regime-tercile analysis shows a trailing, Gaussian risk input has no structural way to anticipate. Cluster stability and regime persistence are the specific evidence a buyer's model-risk function should request before relying on this layer (Table 9).

6.2 Multi-Dimensional Asset Rating

Each tracked asset, more than 450 at present, is scored across market, technical and behavioural dimensions and mapped to a familiar credit-style scale. The intent is not to replace human judgment but to give institutional allocators a common risk vocabulary for an asset class that currently lacks one, refreshed continuously rather than on the periodic cycle of a traditional rating agency. Where an incumbent's risk output is a single portfolio-level VaR or CVaR number (Section 3.5), a continuously updated, asset-level rating gives a CIO or portfolio manager a more granular, forward-looking input to position sizing and mandate compliance, one that moves as an asset's behaviour changes rather than only at the next scheduled review.

6.3 Probabilistic Trend Forecasting

Rather than a single-point price prediction, the platform produces a short-horizon forecast together with an explicit confidence band, so a desk can size a position against a stated degree of uncertainty rather than false precision. This is the capability documented in full in Sections 4 and 5: a five-run ensemble, dual-branch TCN-Transformer architecture, calibrated post hoc with split conformal prediction rather than a Gaussian quantile function, and evaluated out-of-sample, including the point-forecast and directional-accuracy results that qualify its headline calibration figures (Sections 5.3 and 5.6).

It is presented here as one of four integrated capabilities, not the platform's sole output, precisely because a calibrated forecast is most useful to an institutional buyer when it feeds a rating (Section 6.2) and a portfolio optimizer (Section 6.4) rather than standing alone.

6.4 Simulation-Based Portfolio Optimization

Portfolio construction is driven by thousands of simulated market paths calibrated to the fat-tailed, correlation-convergent behaviour described in Section 3, rather than by a single historical covariance matrix, the same static-covariance assumption Section 3.5 identifies as a source of fragility in historical-simulation VaR/CVaR at regime transitions.

The objective function balances downside protection against recovery speed, net of a multi-term diversification penalty combining concentration (Herfindahl-Hirschman Index), entropy (Shannon) and an optional correlation-based Diversification Ratio penalty, complemented by a hard per-asset weight cap. All weighting parameters, including the diversification-penalty intensities, can be recalibrated to a given mandate's risk tolerance.

Out-of-sample drawdown, CVaR, turnover and slippage under this engine are the evidence class Table 9 specifies for enterprise sale, held to the same standard Sections 5.4–5.6 apply to the forecasting engine that feeds it.

6.5 The Trust Layer

Underpinning all four capabilities is a trust layer: risk scores and portfolio decisions are time-stamped and cryptographically verifiable, the same hash-verified, release-manifest discipline applied to the model evaluation reported in Section 5.1, extended here to every behavioural cluster assignment, asset rating and optimizer output the platform produces in production, not only to a published backtest.

For a model-risk or compliance buyer (Table 10), this is what converts "the model said so" into a reproducible, time-stamped decision record that an internal or regulatory reviewer can independently reconstruct.

6.6 Evidence Standards by Layer

Table 9 sets out the specific institutional evidence each layer must produce to support an enterprise sale.

Layer	Institutional problem	BlueHorns capability	Evidence required for enterprise sale
1. Behavioural risk states	False diversification / hidden common exposures across tokens	Unsupervised clustering across market and behavioural features	Cluster stability, regime persistence, out-of-sample validation
2. Asset rating	No common, dynamic risk language across a heterogeneous token universe	Continuous multi-dimensional rating across the covered asset universe	Rating stability, predictive validity, explainability
3. Probabilistic forecasting	Point forecasts encourage false precision (Sec. 5.3)	Forecast distributions with conformal calibration bands (Sec. 4–5)	Calibration curves, pinball/CRPS loss, walk-forward tests
4. Simulation & optimization	Static covariance misses stress-period behaviour (Sec. 3)	Fat-tailed, regime-aware scenario engine + constrained optimization	Out-of-sample drawdown, CVaR, turnover, slippage (Sec. 5.4–5.6)
Trust layer	Model-risk and audit requirements	Time-stamped, hash-verified model/data/decision lineage (Sec. 5.1)	Independent audit, reproducibility, access controls

Table 9. *Product architecture and the institutional evidence standard attached to each layer.*

6.7 Primary Buyer Personas

Buyer	Primary pain	Buying trigger	BlueHorns wedge
CIO / Portfolio Manager	Portfolio construction across many tokens; hidden concentration	Need for repeatable, mandate-compliant allocation	Behavioural clusters + tail-aware optimizer
CRO / Risk Officer	Tail risk, stress testing, independent challenge to vendor VaR	New institutional mandate / regulatory review	Backtested calibration, stress scenarios, risk ratings (Sec. 3–5)
Model Risk / Compliance	Reproducibility and audit trail	Enterprise model approval	Hash-verified decision lineage; full limitations disclosure (Sec. 5.6)
Family office / Private bank	Client suitability and risk communication	Growth in client crypto exposure	Rating + calibrated scenario reporting
Crypto hedge fund	24/7 risk, leverage, derivatives and venue fragmentation	Scaling AUM / reducing internal tooling burden	API + real-time risk + simulation

Table 10. *Primary institutional buyer personas.*

7. Go-to-Market and Commercial Architecture

BlueHorn's go-to-market sells the risk decision, not the dashboard, sequenced across four beachheads:

- **Beachhead 1 — crypto-native hedge funds and digital-asset managers:** shortest sales cycle and strongest pain; integrated through API/PMS workflows.
- **Beachhead 2 — private banks, RIAs and family offices:** standardized risk reports, suitability support and portfolio diagnostics.
- **Beachhead 3 — institutional asset managers and ETF/ETP ecosystems:** model governance, portfolio risk and independent analytics.
- **Beachhead 4 — exchanges, custodians and venues:** risk scoring and portfolio/risk APIs sold as infrastructure.

Geography: United States first, the UAE as a high-value institutional and regulatory hub, then Europe under MiCA and selected Asian financial centres.

Pricing architecture

Package	Indicative ACV	Typical customer	Core deliverable
Professional	$25k–$75k	Small fund / RIA / family office	Ratings + portfolio risk + reports
Institutional	$100k–$250k	Crypto fund / asset manager	API + tail-aware optimizer + scenarios + governance
Enterprise	$250k–$750k+	Bank / exchange / large allocator	Private deployment, SSO, custom models, SLAs, audit

Table 11. *Proposed commercial architecture — an indicative pricing framework to be validated through paid pilots and procurement feedback, not an observed market benchmark.*

8. Conclusion

The crypto portfolio-risk category is real, growing and already contains credible incumbents; the defensible thesis for a new entrant is therefore methodological, not categorical. Sections 3 and 5 make that method explicit and falsifiable: crypto-asset returns reject the normality assumption embedded in classical Gaussian VaR/CVaR, by a margin documented across independent academic studies and independently replicated in this paper on realized BTC-USD outcomes (Section 3); and an alternative built around tail-aware forecasting and distribution-free conformal calibration can be evaluated, including its weaknesses, against exactly the standard a model-risk function would apply to any quantitative claim (Section 5).

BlueHorns AI is an institutional quantitative risk-infrastructure layer for portfolios composed of digital assets. It combines behavioural risk-state detection, continuous asset ratings, calibrated probabilistic forecasts and tail-aware portfolio simulation into an auditable decision system (Section 6) designed for the 24/7, non-Gaussian, regime-sensitive structure of crypto markets, and it publishes the parts of its own validation record that are still in progress alongside the parts that are complete, on the view that this is what makes the complete parts credible.

References

Alexander, C., & Dakos, M. (2023). Assessing the accuracy of exponentially weighted moving average models for value-at-risk and expected shortfall of crypto portfolios. *Quantitative Finance*, 23(3), 393–427.

Amberdata (n.d.). *Risk and Portfolio Management*. Company product documentation.

BCG (2026). *An Imperative for Growth and the New Economics of Asset Management*.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39, 841–862.

Coalition Greenwich / Amberdata (2023). *Digital Assets Infrastructure Study*.

CoinShares (2026a). *Digital Asset Fund Manager Survey — May 2026*.

CoinShares (2026b). *Digital Asset Fund Manager Survey — August 2026*.

Cui, Z., Comerford, D., Crane, M., Nguyen, A., & Bezbradica, M. (2026). Bitcoin Prediction: A Hybrid Approach with Recurrent Neural Networks and GPT-Based Reasoning. In *Artificial Intelligence and Cognitive Science (AICS 2025)*, Communications in Computer and Information Science, vol. 2950. Cham: Springer.

Díaz, A., Esparcia, C., & Huélamo, D. (2023). Stablecoins as a tool to mitigate the downside risk of cryptocurrency portfolios. *North American Journal of Economics and Finance*, 64, 101838.

Gkillas, K., & Katsiampa, P. (2018). An application of extreme value theory to cryptocurrencies. *Economics Letters*, 164, 109–111.

Hasanli, R., & Dursun, M. (2026). A Novel Hybrid Transformer-Based Deep Learning Approach for Multi-Step Bitcoin Price Forecasting. *Engineering, Technology & Applied Science Research*, 16(1), 32524–32533.

Kaiko (n.d.). *Portfolio Risk and Performance*. Company product documentation.

Ko, H., Son, B., & Lee, J. (2024). Portfolio insurance strategy in the cryptocurrency market. *Research in International Business and Finance*, 67, 102135.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3, 73–84.

Kurosaki, T., et al. (2021). Cryptocurrency portfolio optimization with multivariate normal tempered stable processes and Foster-Hart risk. *Finance Research Letters*, 42, 102143.

Li, X., Liu, Z., & Yan, J. (2026). Performance-based regularization for downside-risk cryptocurrency portfolios. *Pacific-Basin Finance Journal*, 103084.

LSTM-conformal forecasting-based Bitcoin forecasting method for enhancing reliability (2025). *PLOS ONE*.

Forecasting bitcoin prices based on hybrid LSTM and deep neural network architectures (2026). *Soft Computing*. Springer.

McKinsey & Company (2025). *Asset management 2025: The great convergence*.

Song, C., et al. (2025). Bitcoin Price Prediction Using Enhanced Transformer and Greed Index. *IET Blockchain*.

State Street (2025). *Digital asset allocations on the rise*.

Subramoney, S. D., Chinhamu, K., & Chifurira, R. (2025). Value at Risk long memory volatility models with heavy-tailed distributions for cryptocurrencies. *Frontiers in Applied Mathematics and Statistics*, 11, 1567626. https://doi.org/10.3389/fams.2025.1567626

Talos (2024). *Acquisition of Cloudwall*. Company announcement.

Talos (n.d.). *Portfolio & Risk Management*. Company product documentation.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. New York: Springer.

Yousefnezhad, P., et al. (2026). Forecasting Economically Significant Bitcoin Moves: A Multi-Scale TCN with Profit-Optimized Thresholds. arXiv:2608.26174.

Zhang, W., Li, Y., Xiong, X., & Wang, P. (2021). Downside risk and the cross-section of cryptocurrency returns. *Journal of Banking & Finance*, 133, 106246.

Appendix A — Statistical Methodology Notes

A.1 Parametric (Gaussian) Value-at-Risk

For a return series assumed $r \sim N(\mu, \sigma^2)$, the p-quantile Value-at-Risk is $VaR_p = \mu + \sigma \cdot \Phi^{-1}(p)$, where Φ^{-1} is the inverse standard-normal CDF. Multi-day horizons are typically obtained by scaling single-day volatility with the square-root-of-time rule, $\sigma_n = \sigma_1 \cdot \sqrt{n}$, which is exact only under i.i.d. Gaussian returns.

A.2 Kupiec unconditional-coverage (POF) test

Given n observations and x breaches of a VaR band with nominal breach probability p, the Kupiec (1995) likelihood-ratio statistic is $LR = -2 \cdot \ln[(1-p)^{n-x} \cdot p^x] + 2 \cdot \ln[(1-\hat{y})^{n-x} \cdot \hat{y}^x]$, where $\hat{y} = x/n$ is the empirical breach rate. Under the null hypothesis that the model is correctly calibrated, LR is asymptotically χ^2 with one degree of freedom; a p-value below 0.05 rejects correct calibration. Section 3.4 applies this test to both the naive Gaussian VaR band and BlueHorn's conformal interval on the same 512-day sample.

A.3 Split conformal prediction

Split conformal prediction (Vovk, Gammerman & Shafer, 2005) calibrates an interval without assuming a parametric error distribution. A non-conformity score — here, max(q_low − y, y − q_high, 0), the amount by which a validation-set outcome y falls outside the model's raw quantile interval [q_low, q_high] — is computed on a held-out calibration (validation) set. The $(1-\alpha)$-quantile of that score, δ, is then added to the upper bound and subtracted from the lower bound of the model's interval on new, unseen data. Under the assumption that calibration and test residuals are exchangeable, the resulting interval has a finite-sample marginal coverage guarantee of at least $1-\alpha$ — without any assumption that the underlying return distribution is Gaussian, heavy-tailed, or any other parametric family.

A.4 Distributional diagnostics

- **Skewness** measures asymmetry; zero for a symmetric distribution such as the Gaussian. Negative skewness indicates a heavier left (downside) tail.
- **Excess kurtosis** measures tail weight relative to the Gaussian, for which excess kurtosis is defined as zero. Positive excess kurtosis (“leptokurtosis”) indicates more probability mass in the extreme tails than a normal distribution predicts.

- **Jarque-Bera test** is a joint test of the null hypothesis that a sample's skewness and excess kurtosis are both zero (i.e., consistent with normality), based on a statistic that is asymptotically χ^2 with two degrees of freedom.
- **Shapiro-Wilk test** is a separate, generally more powerful, small-to-moderate-sample test of the null hypothesis that a sample is drawn from a normal distribution.

A methodological note on Section 3.4: the 3-day realized log-returns used there are constructed on overlapping daily windows, which induces serial correlation the Jarque-Bera and Shapiro-Wilk tests do not model. This affects the exact size of the reported p-values but not the qualitative conclusion — both the magnitude of the estimated excess kurtosis (4.08) and its consistency with the non-overlapping daily-return statistics surveyed from the literature in Section 3.2 support the same finding independently of this effect.