Forecasting Black Swan Events: Financial Markets Volatility Analysis & Extreme Event Prediction

BOUAOUICHE Haidar, DAS Modhura, LUO Wei, O'BRIEN Gerry, REBUT Quentin

December 23, 2025

Abstract

Financial models frequently assume asset returns follow a normal distribution, simplifying risk management but dangerously underestimating the probability of extreme market shocks or "Black Swans". This project challenges the Gaussian assumption by empirically analyzing S&P 500 returns from 2005 to 2025. We demonstrate that financial markets exhibit fat tails and volatility clustering, phenomena that classical models fail to capture. By implementing a GARCH(1,1) model with Student's t innovations, we develop a dynamic "Early Warning System" that forecasts time-varying tail risk. Our results show that while Black Swan events cannot be precisely timed, their statistical fragility can be detected, aligning with industry standards for robust risk management.

1 Introduction & The Gaussian Illusion

Financial theory has long relied on the assumption that asset returns are normally distributed. This Gaussian framework, while computationally convenient, implies that extreme events are astronomically rare. A "Black Swan" event, a term popularized by Nassim Taleb, describes highly improbable, high-impact occurrences that standard models deem effectively impossible. However, history (e.g., the 2008 Financial Crisis, the 2020 COVID-19 crash) demonstrates that these events occur with consequential regularity.

The core tension of our study is the mismatch between classical thin-tailed models and the fat-tailed reality of financial markets. In the "Fourth Quadrant", the domain of complex, payoff-dominated events, relying on the normal distribution is not just inaccurate; it is dangerous. Our goal is not to predict the exact timing of a crash but to identify periods of statistical fragility where the risk of a Black Swan is elevated.

2 Diagnostic Phase: Looking Past Normality

We analyzed daily S&P 500 prices from 2005 through early 2025, computing daily log returns. This dataset captures multiple structural breaks and systemic failures, providing a rich testing ground for extreme event modeling.

2.1 Evidence of Non-Normality

Our diagnostic tests unequivocally rejected the normality assumption.

- **Visual Evidence:** The histogram of returns shows a "leptokurtic" shape, a higher peak and fatter tails than the normal curve. The Q-Q plot reveals significant deviation from the diagonal in the tails, indicating that extreme gains and losses occur far more frequently than predicted by a Gaussian model. Under normality, the empirical quantiles should align

closely with the theoretical normal quantiles, forming a straight diagonal line. Instead, we observe very clear misalignment in both tails, confirming that large market movements are not statistical outliers but recurring features of the data.

- **Statistical Tests:** Both the Kolmogorov-Smirnov ($p \approx 0$) and Anderson-Darling tests strongly rejected the null hypothesis of normality. This statistical rejection implies that models assuming Gaussian returns will systematically underestimate the frequency and severity of extreme events, precisely the conditions under which Black Swans emerge.

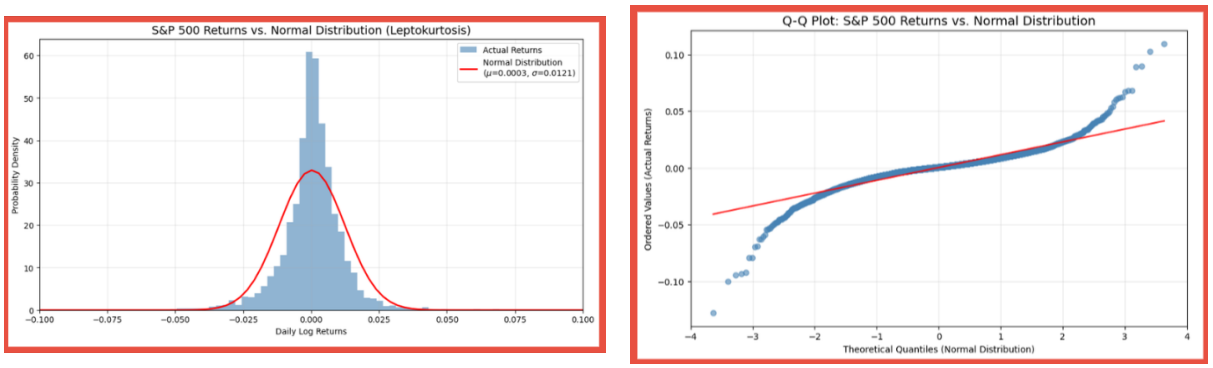

Figure 1: S&P 500 Returns vs. Normal Distribution. The Histogram (a) shows a higher peak and fatter tails than the normal curve, while the Q-Q plot (b) demonstrates severe tail deviation.

These findings confirm that S&P 500 returns are neither independent nor identically distributed (i.i.d.) normal variables, necessitating a modeling approach that accounts for heavy tails.

3 Proof of Volatility Clustering

Before employing GARCH-type models to forecast volatility and extreme events in financial markets, it is imperative to establish the presence of **volatility clustering** in the data. Volatility clustering is a well-documented empirical regularity in financial time series, first put forward by Mandelbrot (1963), whose words were "*large changes tend to be followed by large changes, and small changes tend to be followed by small changes.*"

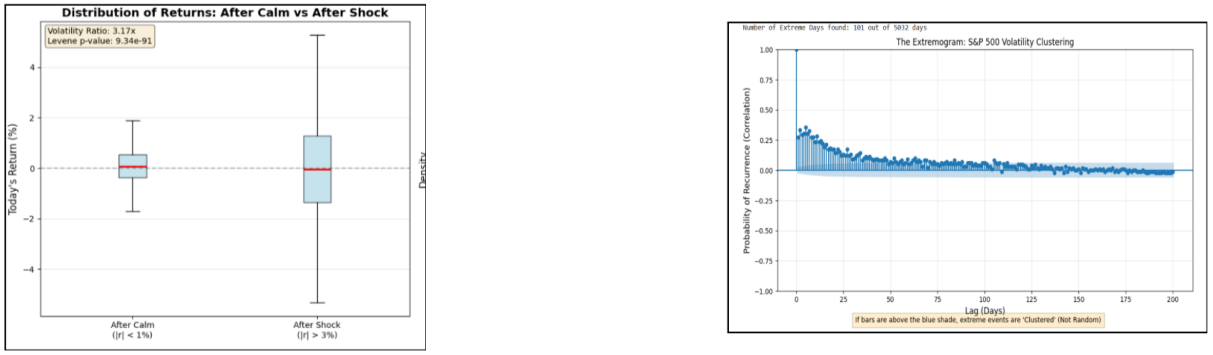

Figure 2: Visual evidence of **volatility clustering** : Clear dependence of future return variability on past return magnitude

irst, we analyzed next-day return distributions conditional on prior market movements. Following **calm periods** ($|r| < 1\%$), returns remain confined to a narrow range ($\approx \pm 2\%$), indicating low volatility. In contrast, **shock periods** ($|r| > 3\%$) are followed by significantly higher variance, with fat tails extending beyond $\pm 5\%$.

Second, the **extremogram**, an autocorrelation function for tail events, quantifies the probability of recurring extremes. Our results show statistically significant dependence lasting 50–100 trading days. This persistence decisively rejects the i.i.d. hypothesis, confirming that extreme events cluster in time and necessitating dynamic volatility modeling.
Therefore, these findings strongly justify our future step using the GARCH-family models.

4 The Solution: GARCH(1,1) with Student's t Innovations

To model this clustering and fat-tailed behavior, we chose a GARCH(1,1) specification with Student's t distributed errors.
The model is defined by two coupled equations: the *Mean Equation*, which describes the returns, and the *Variance Equation*, which describes the evolution of volatility.

$$r_t = \mu + \epsilon_t, \quad \text{where } \epsilon_t = \sigma_t z_t, \quad z_t \sim t_\nu(0,1) \tag{1}$$
$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{2}$$

Here, r_t is the daily log return, μ is the constant mean return, and ϵ_t is the residual shock. The innovation z_t follows a Student's t distribution with ν degrees of freedom, accounting for the heavy tails identified in our diagnostic phase.

4.1 Why GARCH(1,1)?

We deliberately chose the parsimonious (1,1) lag structure, one lag for the shock (α) and one lag for the persistent variance (β), rather than a higher-order model (e.g., GARCH(2,2)).

- **Parsimony Principle:** Empirical research (e.g., Hansen & Lunde, 2005) consistently demonstrates that a simple GARCH(1,1) model outperforms more complex specifications in forecasting daily equity volatility.
- **Avoidance of Overfitting:** Higher-order models often introduce statistically insignificant parameters and instability without improving out-of-sample predictive power. The (1,1) specification is sufficiently flexible to capture both the immediate reaction to news and long-term volatility persistence.

4.1.1 Estimated Parameters & Interpretation

Our estimated parameters confirm the persistent and fragile nature of market volatility:

- $\alpha \approx 0.13$: The **reaction** parameter. This indicates that roughly 13% of a variance shock is immediately priced into the next day's volatility.
- $\beta \approx 0.86$: The **persistence** or "memory" parameter. This shows that 86% of yesterday's variance carries over to today, explaining why calm periods follow calm periods.
- $\alpha + \beta \approx 0.99$: The **persistence sum**. A value close to 1 implies that volatility shocks decay extremely slowly. Once a crisis hits, the market remains in a high-volatility regime for weeks or months (long memory).
- Degrees of Freedom ($\nu \approx 5.3$): This parameter controls the shape of the tail. A normal distribution corresponds to $\nu = \infty$. A value as low as 5.3 confirms **extremely fat tails**, meaning 3-sigma or 4-sigma events are far more likely than in a Gaussian world.

4.1.2 Industry Application

In the financial industry, this specific GARCH framework is a cornerstone of quantitative risk management. It is widely used for:

- **Value-at-Risk (VaR) Estimation:** Banks use GARCH-based VaR models to calculate regulatory capital requirements under Basel III standards.
- **Volatility Targeting:** Asset managers use dynamic volatility forecasts to scale position sizes (leveraging up in calm periods, deleveraging in crises).

4.2 Sanity Check: Comparison with VIX

To validate our model, we compared our estimated conditional volatility with the VIX Index (the market's forward-looking "fear gauge") and Realized Volatility. The GARCH volatility tracks these benchmarks almost perfectly, capturing every major crisis (2008, 2011, 2020). This confirms that our statistical model is successfully extracting the latent market risk signal.

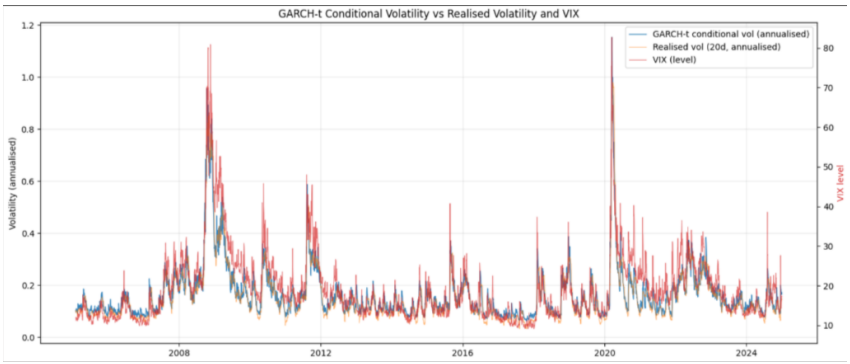

Figure 3: GARCH-t volatility aligns with the VIX Index, validating its ability to capture market fear regimes.

5 From Volatility to Tail Risk: The Early Warning System

Moving beyond abstract volatility numbers, we constructed an actionable metric for risk management: the **Dynamic Crash Probability**. We define a "crash" as a daily log return dropping below a threshold $c = -5\%$.

5.1 Methodology

Using the conditional volatility forecasts σ_{t+1} generated by our GARCH-t model, we calculate the probability that tomorrow's return r_{t+1} will breach this threshold, conditional on today's information $\mathcal{F}_t$:

$$P(r_{t+1} < -5\% | \mathcal{F}_t) = P\left(z_t < \frac{-0.05 - \mu}{\sigma_{t+1}}\right) \tag{3}$$

where z_t follows a standardized Student's t distribution with degrees of freedom ν. This formula effectively standardizes the crash threshold by the current market volatility, allowing us to read the probability directly from the heavy-tailed distribution.

5.2 Distributional Analysis

Applying this model to the full dataset reveals a realistic profile of market risk:

- **Base Rates:** The median crash probability is extremely low at 0.06%, confirming that on most days, the market is stable. The upper quartile is only 0.32%, meaning that 75% of the time, the risk of a crash is negligible.
- **Crisis Spikes:** The average probability is pulled up to 0.66% by extreme spikes during crisis periods.
- **Reality Check:** In the actual historical data, we observed roughly 20 days with losses worse than -5% out of $\sim$5,000 observations (a base rate of $\sim 0.4\%$). Our model's probabilities align well with this reality, effectively filtering out noise while reacting aggressively to true danger signals.

5.3 Performance Validation

To validate the model's utility as an Early Warning System, we analyzed its performance on specific crash days versus non-crash days:

- **Signal Clarity:** On days where a crash actually occurred, the model's average implied probability was approximately **12%**. In contrast, on non-crash days, it was less than **1%**.
- **Risk Capture:** Almost all realized crash days were captured within the top 10% of the model's most "risky" days.
- **Correlation with Benchmarks:** We compared the Dynamic Crash Probability with the VIX Index and Realized Volatility. The probability metric drops to near zero during calm periods, unlike the VIX, which often maintains a baseline level, suggesting our metric is more finely tuned specifically to *crisis* risk rather than general sentiment.

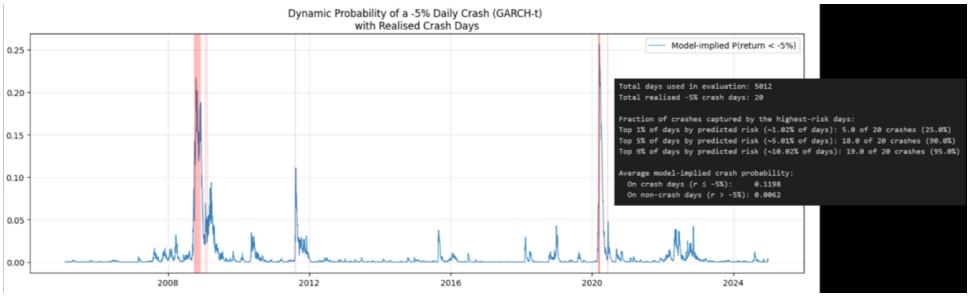

Figure 4: Dynamic Crash Probability (2005-2025). The model assigns high probability spikes (y-axis) that align perfectly with actual market crashes (red markers), specifically in 2008, 2011, and 2020.

6 Out-of-Sample Testing: The 2019-2024 Challenge

To rigorously test predictive power, we performed a static Out-of-Sample (OOS) test. We split the data into a **Training Set (2004–2018)** and a **Test Set (2019–2024)**. This design ensures the model "learns" from the 2008 Global Financial Crisis but must blindly navigate the 2020 COVID-19 crash.

6.1 Training Phase (2004–2018)

The model estimated on the training set revealed extreme volatility persistence. The sum of lag parameters was $\alpha + \beta \approx 0.9988$. This near-unity persistence implies that volatility shocks decay incredibly slowly; a shock today affects risk forecasts months into the future. The degrees of freedom $\nu \approx 5.29$ confirmed the presence of heavy tails in the training data.

6.2 Test Results (2019–2024)

Applying these fixed pre-2018 parameters to the unseen 2019-2024 data yielded robust results:

- **Regime Identification:** The model correctly identified the March 2020 COVID-19 crash cluster as the highest-risk period in the test set.
- **Capture Rate:** In the out-of-sample period, 67% of all actual crashes occurred within the days our model flagged as "High Risk" (the top 4.6% of probability scores).

This demonstrates that while the model cannot predict the *event*, it successfully identified the resulting *fragility* and high-volatility regime that made the market susceptible to extreme moves.

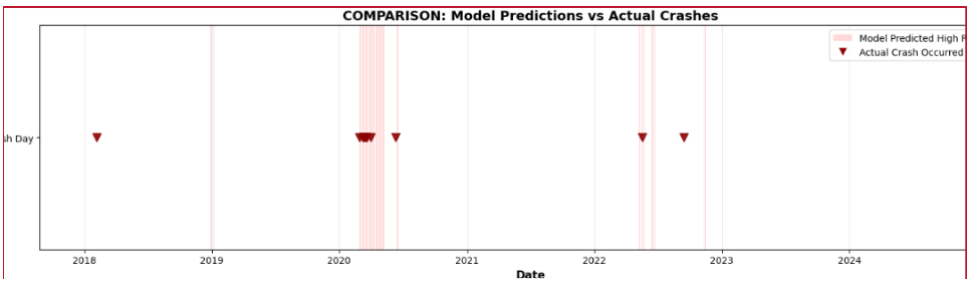

Figure 5: Out-of-Sample Performance (2019-2024). Despite being trained only on pre-2018 data, the model successfully signaled extreme risk during the 2020 COVID crisis.

7 Conclusion & Future Extensions

7.1 The Rolling Window Extension

While our static split validated the model's mechanics, financial markets are non-stationary, structural breaks occur that can render old parameters obsolete. To align with industry standards, the logical next step is a **Rolling Window Methodology**.
By re-calibrating the model daily using a fixed window (e.g., the last 1,000 days), the system adapts to new volatility regimes in real-time. This eliminates look-ahead bias and simulates the constraints faced by a risk manager. Although computationally intensive, this extension transforms the project from a statistical study into a professional-grade Early Warning System.

7.2 Final Verdict

Our project aimed to verify if Black Swan events, despite their definition as "unpredictable," leave statistical footprints. We conclude that:

1. **The Gaussian Illusion is Dangerous**: Normal distribution models fundamentally misprice risk in the "Fourth Quadrant."
2. **Volatility is Predictable**: Clustering is a statistically significant phenomenon that can be modeled with GARCH.
3. **Fragility can be Forecasted**: While we cannot predict the specific timing or trigger of a Black Swan, our GARCH-t model successfully identifies periods of elevated fragility.

As Nassim Taleb argues, the goal in the Fourth Quadrant is not to build a perfect prediction machine, but to build robustness against tails. Our model serves this purpose by providing a reliable signal of when to play defense.

A Appendix

A.1 Daily Log Returns - S&P500

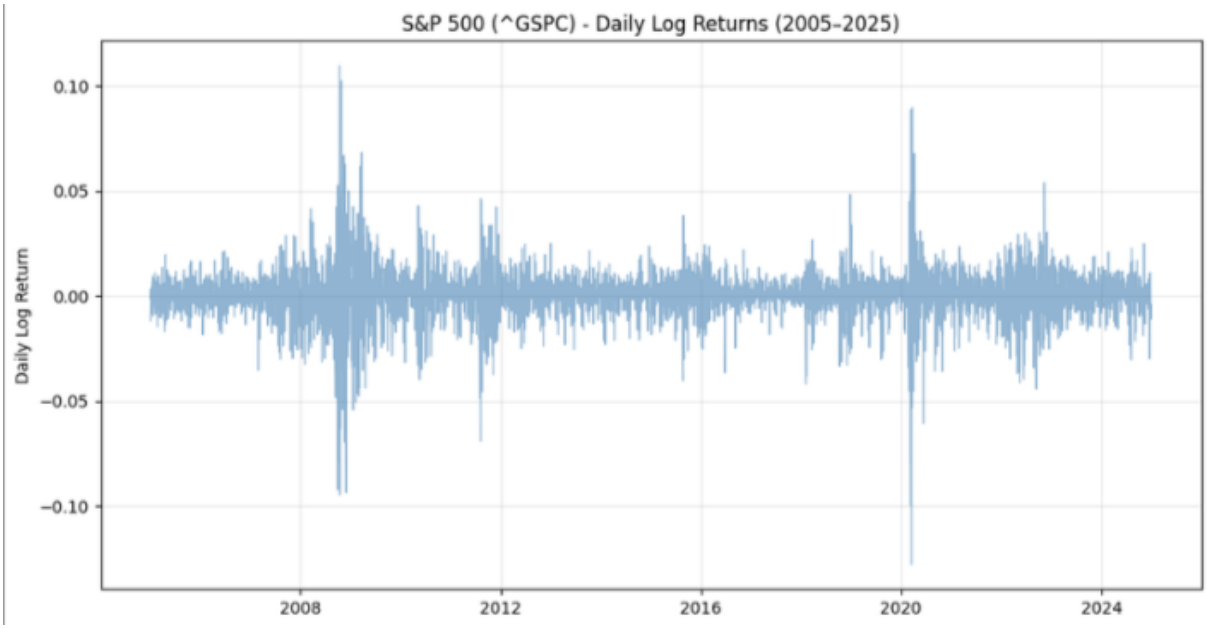

Figure 6: Visual Representation of our original datasat : Daily Log Returns of S&P500 between 2005 and 2025, from *yfinance* library in Python

A.2 Why Student's t ?

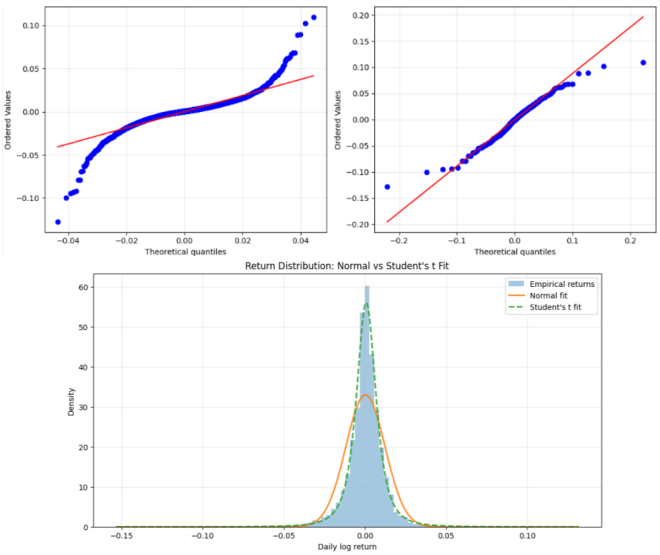

Figure 7: Comparative Distributional Fit. The **Q-Q Plots** (top) demonstrate that the Student's t-distribution aligns with the data's tails far better than the Normal distribution. The **Histogram** (bottom) confirms this, showing that the Student's t curve (dotted green) captures the heavy tails and higher peak of S&P 500 returns, whereas the Gaussian curve (orange) underestimates extreme frequencies.

A.3 Correlation with external indicators of fear and turbulence

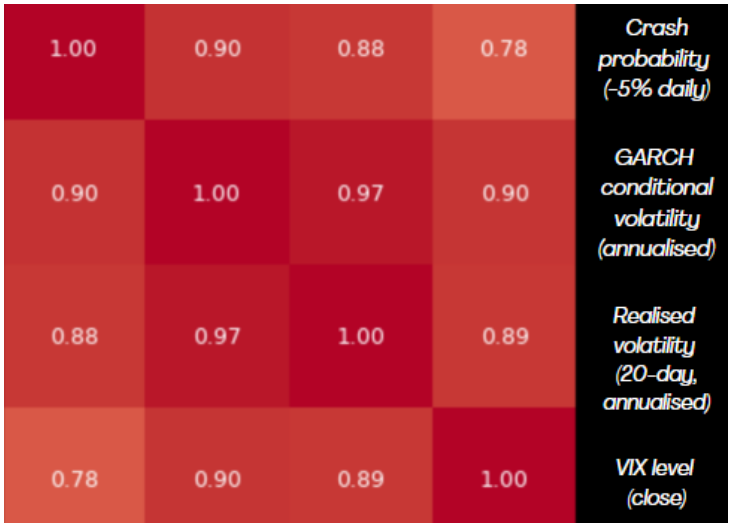

Figure 8: Correlation between Model Outputs and Market Benchmarks. All correlations are significant at $p < 0.01$.

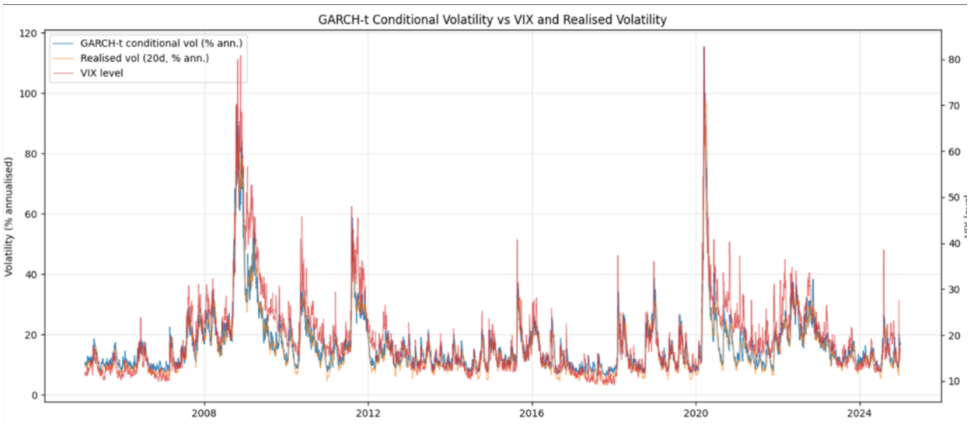

Figure 9: GARCH-t Conditional Volatility vs Market Benchmarks

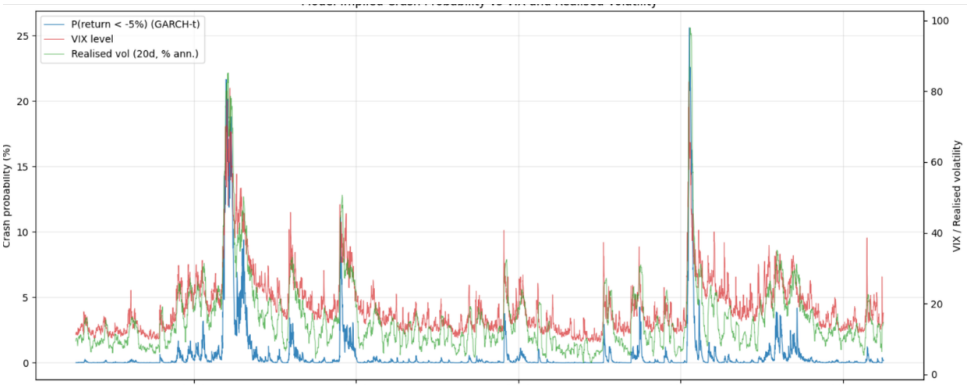

Figure 10: Probability of Crash (return $< -5\%$) vs Market Benchmarks

Validation against Market Benchmarks: The strong positive correlation between the model-implied Crash Probability and the VIX indicates that the GARCH-t model effectively

captures the same high-risk regimes priced into the options market.

A.4 Model Estimation & Out-of-Sample Performance

This appendix details the specific parameters estimated during the Training Phase (2004–2018) and the quantitative performance metrics observed during the Testing Phase (2019–2024).

A.4.1 Estimated GARCH(1,1)-t Parameters

The model was fitted on daily S&P 500 log returns from Jan 2004 to Dec 2018. The resulting parameters (Table 1) define the volatility dynamics used for the out-of-sample forecast.

Parameter	**Coefficient**	**Std. Error**	***t*-statistic**	**P-value**
Mean Equation				
μ (Constant)	0.0736	0.0114	6.437	< 0.001
Variance Equation				
ω (Base Variance)	0.0109	0.0038	2.838	0.042
α_1 (Reaction)	0.1143	0.0166	6.867	< 0.001
β_1 (Persistence)	0.8845	0.0162	54.685	< 0.001
Distribution				
ν (Degrees of Freedom)	5.2912	0.5170	10.231	< 0.001

Table 1: GARCH(1,1) Estimation Results (Training Set: 2004–2018). All key parameters are statistically significant at the 1% level.

Interpretation of Results:

- **Persistence ($\alpha + \beta$):** The sum is $0.1143 + 0.8845 = 0.9988$. This is exceptionally close to 1, indicating an Integrated GARCH (IGARCH) process where volatility shocks have a near-permanent effect on future risk forecasts.
- **Fat Tails (ν):** The estimated degrees of freedom is 5.29, significantly lower than the normal distribution (∞). This confirms that the model anticipates extreme outliers as a structural feature of the market.

A.4.2 Predictive Power (2019–2024)

To assess the model's utility as an Early Warning System, we analyzed its ability to flag crash days (returns $< -4\%$) in the unseen test set. The threshold of a crash has been chosen in order to have a larger set of positive cases.

Metric	Value
Total Crash Days (Test Set)	12
Crashes Correctly Flagged	8
Capture Rate	**67%**
COVID CRASH PERIOD (Feb-Apr 2020)	
Average crash probability	13.03%
Max crash probability	29.10%
Actual crash day	8

Table 2: Out-of-Sample Performance Metrics. The model successfully captured 2/3rds of all crashes while flagging only a small fraction of days as high risk.

The capture rate of 67% (i.e. 67% of all crashes occurring within the 4.6% of days flagged as high-risk) confirms that the model identified the fragile regime preceding the March 2020 COVID-19 crash, validating the efficacy of the GARCH-t specification for tail risk forecasting.

A.5 Some references to Taleb's work

A.5.1 Theoretical Framework : Detecting Fragility

Our statistical findings align closely with the risk philosophy established by Nassim Taleb. While standard econometric models often treat extreme events as statistical anomalies to be smoothed out, our project operates in what Taleb defines as the "Fourth Quadrant", the domain of complex, payoff-dominated events where "Black Swans" reside.

A.5.2 Prediction vs. Fragility Detection

A common criticism of tail-risk modeling is the impossibility of predicting specific Black Swan events. However, as noted in the slide analysis:

> *"You cannot predict Black Swans, but you can detect fragility."*
> Nassim Taleb

Our GARCH-t model validates this concept. It does not predict *what* will cause a crash (e.g., a virus or a bank failure), but it successfully identifies *when* the market has entered a state of susceptibility. As Taleb argues in *Antifragile* (2012): *"After a large shock, systems remain in a state of fragility for a long time."*. This is empirically demonstrated by our model's high persistence parameter ($\alpha + \beta \approx 0.998$), showing that volatility clusters and shocks do not dissipate quickly.

A.5.3 The Danger of Thin Tails

The visual comparison below (Figure 11) highlights the danger of relying on Gaussian assumptions.

- **Top Panel (Normal):** The GARCH model with Normal errors (yellow) creates a false sense of security. It barely registers the COVID-19 crash probability above the 5% threshold, essentially "underestimating extreme risk".
- **Bottom Panel (Student's t):** The GARCH model with Student's t errors (blue) breaches the 5% crash-probability threshold frequently during crises. This is not a flaw, but a **strength**. It correctly signals that the system is fragile, enabling proactive rather than reactive risk management.

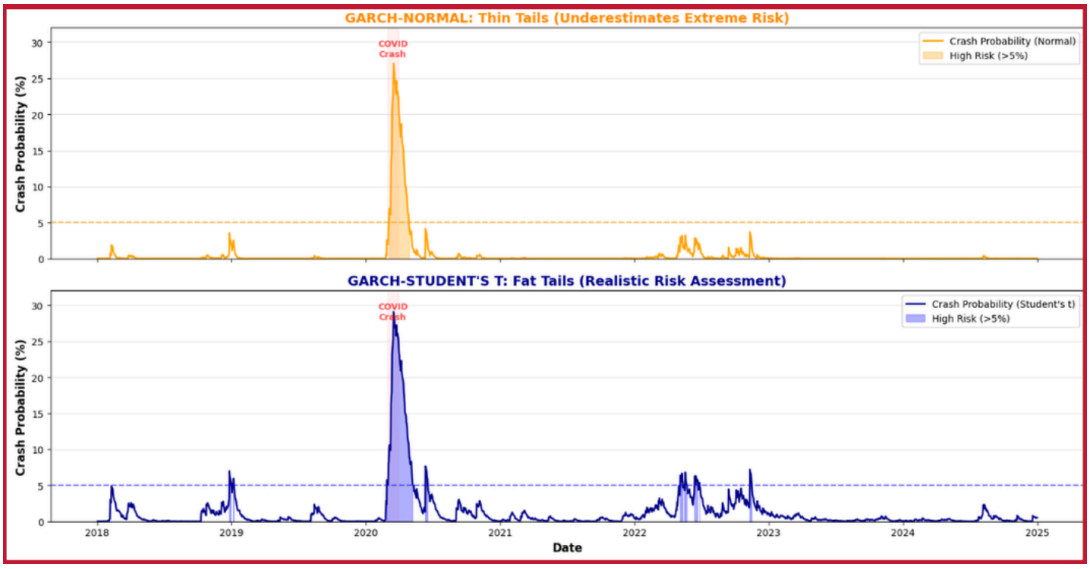

Figure 11: Visualizing Fragility. The Normal model (top) suppresses tail risk, while the Student's t model (bottom) accurately reflects the "fat tails" and fragility clusters described by Taleb.

A.5.4 Conclusion on Modeling Philosophy

Our results confirm that *"Rare events cannot be treated as outliers; they are the main drivers of history"* (Taleb, 2007). By using heavy-tailed distributions ($v \approx 5.3$), our Early Warning System adheres to the principle that while all models are wrong, those that account for fat tails are "less wrong" and significantly safer for risk management than their Gaussian counterparts.

References

[1] Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2), 223–236.

[2] Cont, R. (2007). Volatility Clustering in Financial Markets: Empirical Facts and Agent-Based Models. In *Long Memory in Economics*. Springer.

[3] Guo, H., & Philippatos, G. C. (2017). Forecasting the volatility of S&P 500 index with GARCH-type models using Student's t-distribution. *Journal of Forecasting*, 36(6), 656–667.

[4] Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36(4), 394–419.

[5] McNeil, A. J., & Frey, R. (2000). Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: An Extreme Value Approach. *Journal of Empirical Finance*, 7(3), 271–300.

[6] McNeil, A. J., Frey, R., & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.

[7] Zivot, E. (2009). Practical Issues in the Analysis of Univariate GARCH Models. In *Handbook of Financial Time Series* (pp. 113–155). Springer-Verlag.