Peer-Review

Chen, Charles. 2026. “Beyond Normality: Comparative Tail-Risk Analysis of S&P 500 Returns.” *Journal of High School Science* 10 (2): 355–73. https://doi.org/10.64336/001c.162826.

1.The central claim—S&P 500 returns are non-Gaussian with fat tails and negative skewness—is not new. It has been known for decades. You have essentially replicated facts, which is not a novel contribution.“Normality is violated” is not publishable unless:

•You show new conditions where it breaks differently, or

•You quantify practical consequences beyond what is already known, or

•You introduce methodological improvement over standard models

The entire motivation of the paper is that Normality underestimates downside risk, so the natural question is Does the GMM produce more accurate tail risk estimates than a normal model? With rolling windows, does the model better represent the true distribution going forward? Can the model predict extreme event frequency (Probability of returns $< -3\sigma$ (or similar thresholds; this directly ties to the fat tail claim)

2.For your VaR section as well, that a normal distribution underestimates risk is well known. You need to provide backtesting of VaR violations , comparison to industry-standard alternatives (e.g., EVT, GARCH, t-distributions) and calculate economic impact (e.g., capital misallocation, drawdown risk).

3.Fat tails and extreme values are not noise, not measurement error. They are structural features of financial returns. So any method - including isolation forest- that removes or compresses extremes is altering the phenomenon rather than analyzing it. This section answers the wrong question: Instead of “Are returns normal after cleaning?”, the relevant question is: “How should models account for extreme events that are structurally embedded?”This creates a logical contradiction: You remove extremes → distribution looks more normal → conclude “less non-normality”. That is tautological, not scientific. If the goal is risk and distribution modeling, better approaches are:

A. Model the tails explicitly

Extreme Value Theory (EVT)

Peaks-over-threshold (POT)

B. Use heavy-tailed distributions

Student-t

skewed-t

C. Model time structure

GARCH / stochastic volatility

D. Evaluate consequences

VaR backtesting (Kupiec test, Christoffersen test)

E. Link fat tails to volatility clustering, macro regimes

and/or crisis periods

•Since tail behavior is central to the paper’s research question (non-normality and VaR underestimation), truncating or clipping extremes biases the analysis and undermines the conclusions. A more appropriate approach would model tail behavior explicitly rather than suppress it.

4.The Gaussian Mixture Model is used primarily as a heuristic curve-fitting device to better match the empirical distribution, but the manuscript does not provide model selection criteria (e.g., AIC/BIC), likelihood-based comparisons, or out-of-sample validation to justify its use. As such, the improvement over a single Gaussian appears descriptive rather than analytically meaningful. A GMM fit alone is not meaningful unless It improves forecasting, It improves risk estimation and It captures economic regimes with generalization. None of these are demonstrated.

5.With n = 7804 daily observations, any deviation will be statistically significant. The rejection of normality is to be expected and is hence not informative. You are missing effect size interpretation, robustness to subsampling and time-varying structure (non-stationarity).

6.You have a serious modeling limitation in that there is no temporal structure modeling. Financial returns are heteroskedastic (volatility clustering), regime-dependent and time-dependent, yet the paper treats returns as i.i.d. samples from a static distribution.
7.No Out-of-Sample Validation. All analyses are descriptive and In-sample. There is no predictive validation, rolling-window analysis or regime sensitivity. None of the models are tested for usefulness.
8.The hypothesis "SPY returns are not perfectly normally distributed" is already established and trivially true. A stronger framing is required. Eg.
•How much does non-normality matter for risk estimation?
•Which models best capture tail risk?
•Under what conditions does normality approximation fail?
9.The GMM is evaluated only by Visual fit (histograms, KDE overlays) and In-sample description
It does not answer:
Does GMM improve VaR accuracy?
Does it reduce tail underestimation?
Does it generalize out-of-sample?
The paper is entirely descriptive:"Here is what the distribution looks like". But financial modeling requires predictive validity and decision relevance
Without prediction, the model is just a smoother curve, not a better model. The paper never demonstrates that its "better-fitting" model improves prediction of any of these quantities (Tail risk (VaR/ES, full return distribution and volatility/extremes).
The manuscript evaluates model performance purely in terms of in-sample distributional fit (e.g., visual agreement with histograms and Q-Q plots). However, it is unclear what quantity the model is intended to predict. If the goal is improved risk estimation, the analysis should evaluate out-of-sample predictive performance, such as VaR exceedance rates or likelihood-based scoring under a rolling-window framework. Without such evaluation, the improved fit of the Gaussian Mixture Model remains descriptive rather than practically meaningful.

---------- General Response ----------

Based on the review comments, the revised manuscript underwent a substantial restructuring. The original emphasis on normality testing was significantly reduced, and the paper now focuses more broadly on how different distributional models affect tail-risk estimation and downside-risk prediction under varying market conditions and confidence levels. In addition, the previous outlier-removal discussion was replaced with a sensitivity analysis.
I acknowledge that several reviewer suggestions involved advanced quantitative finance techniques that would substantially expand the scope of the study. I carefully reviewed and incorporated indirectly through expanded rolling-window analysis, exceedance backtesting, yearly regime comparison, sensitivity testing, and comparative tail-risk modeling.
I also recognize that the non-normality of financial returns is already well documented in the financial literature. Rather than focusing solely on rejecting normality, the revised paper systematically compares how different distributional models affect practical tail-risk estimation, model calibration, and downside-risk prediction under varying market regimes and tail depths, and emphasizes the practical consequences of distributional model selection for risk management.
Because of the significant changes in the paper's focus and structure, I changed the paper title from "A Comparative Analysis of S&P 500 Return Normality Using Statistical and Machine Learning Methods" to "Beyond Normality: Comparative Tail-Risk Analysis of S&P 500 Returns."

---------- Responses to each individual comment ----------

1. The central claim—S&P 500 returns are non-Gaussian with fat tails and negative skewness—is not new. It has been known for decades. You have essentially replicated facts, which is not a novel contribution. "Normality is violated" is not publishable unless: You show new conditions where it breaks differently, or You quantify practical consequences beyond what is already known, or You

introduce methodological improvement over standard models The entire motivation of the paper is that Normality underestimates downside risk, so the natural question is Does the GMM produce more accurate tail risk estimates than a normal model? With rolling windows, does the model better represent the true distribution going forward? Can the model predict extreme event frequency (Probability of returns $< -3\sigma$ (or similar thresholds; this directly ties to the fat tail claim)

Response: I agree with the reviewer that the non-normality of financial returns is already well documented in financial literature. As discussed in the General Response above, the revised manuscript was substantially refocused from a simple normality-testing study toward practical tail-risk estimation and model calibration under different market conditions.

To address these concerns, the revised paper now includes comparative VaR and ES analysis, exceedance backtesting, rolling-window analysis, annual regime-based calibration analysis, sensitivity analysis of extreme observations, and AIC/BIC model comparison. The revised analyses evaluate not only whether normal distribution fail, but also under what conditions they fail and which alternative models better capture tail-risk behavior and extreme-loss frequency.

The revised paper concludes that no single model universally dominates across all market conditions and tail depths, with model performance depending strongly on market regime, confidence level, and forecasting objective.

2. For your VaR section as well, that a normal distribution underestimates risk is well known. You need to provide backtesting of VaR violations , comparison to industry-standard alternatives (e.g., EVT, GARCH, t-distributions) and calculate economic impact (e.g., capital misallocation, drawdown risk).

Response:: As discussed in the General Response, the revised paper now includes empirical historical VaR, Normal VaR, t-distribution VaR, two-component GMM VaR, EVT-based VaR, and Expected Shortfall (ES) comparisons across multiple confidence levels.

Most importantly, VaR exceedance backtesting and 500-day rolling-window analysis were added to evaluate dynamic predictive performance and actual exceedance frequency under changing market conditions

However, advanced approach such as GARCH modeling was beyond the intended scope of the current study.

3. Fat tails and extreme values are not noise, not measurement error. They are structural features of financial returns. So any method - including isolation forest- that removes or compresses extremes is altering the phenomenon rather than analyzing it. This section answers the wrong question: Instead of "Are returns normal after cleaning?", the relevant question is: "How should models account for extreme events that are structurally embedded?"This creates a logical contradiction: You remove extremes →distribution looks more normal →conclude "less non-normality". That is tautological, not scientific. If the goal is risk and distribution modeling, better approaches are: A. Model the tails explicitly Extreme Value Theory (EVT) Peaks-over-threshold (POT) B. Use heavy-tailed distributions Student-t skewed-t C. Model time structure GARCH / stochastic volatility D. Evaluate consequences VaR backtesting (Kupiec test, Christoffersen test) E. Link fat tails to volatility clustering, macro regimes and/or crisis periods Since tail behavior is central to the paper's research question (non-normality and VaR underestimation), truncating or clipping extremes biases the analysis and undermines the conclusions. A more appropriate approach would model tail behavior explicitly rather than suppress it.

Response: I agree with the reviewer that extreme observations are structural features of financial returns rather than simple noise. As discussed in the General Response, the original outlier-removal discussion was removed from the revised manuscript and replaced with a limited sensitivity analysis evaluating model robustness to rare extreme events.

4. The Gaussian Mixture Model is used primarily as a heuristic curve-fitting device to better match the empirical distribution, but the manuscript does not provide model selection criteria (e.g., AIC/BIC), likelihood-based comparisons, or out-of-sample validation to justify its use. As such, the improvement over a single Gaussian appears descriptive rather than analytically meaningful. A

GMM fit alone is not meaningful unless It improves forecasting, It improves risk estimation and It captures economic regimes with generalization. None of these are demonstrated.
Response: As discussed in the General Response, the revised manuscript expanded the analysis beyond descriptive distribution fitting. Formal model-selection analysis using log-likelihood, AIC, and BIC was added to directly compare the normal distribution, t-distribution, and two-component GMM models while accounting for model complexity.
In addition, the revised manuscript now evaluates two-component GMM using VaR estimation and exceedance backtesting rather than relying solely on histogram fitting. The revised results show that although GMM improves fit relative to the single Gaussian model, the t-distribution achieves the strongest overall balance between model fit, complexity, and tail-risk representation among the tested parametric models.
5. With n = 7804 daily observations, any deviation will be statistically significant. The rejection of normality is to be expected and is hence not informative. You are missing effect size interpretation, robustness to subsampling and time-varying structure (non-stationarity).
Response: I agree that with very large sample sizes, formal rejection of normality alone becomes less informative. As discussed in the General Response, the revised manuscript reduced emphasis on statistical rejection itself and instead focused more heavily on practical tail-risk estimation, model calibration, and predictive performance under different market conditions.
6. You have a serious modeling limitation in that there is no temporal structure modeling. Financial returns are heteroskedastic (volatility clustering), regime-dependent and time-dependent, yet the paper treats returns as i.i.d. samples from a static distribution.
Response: As discussed in the General Response, the revised manuscript now includes rolling-window backtesting, annual regime-based analysis, and exceedance-frequency evaluation to partially address time-varying market behavior and dynamic tail-risk estimation.
7. No Out-of-Sample Validation. All analyses are descriptive and In-sample. There is no predictive validation, rolling-window analysis or regime sensitivity. None of the models are tested for usefulness.
Response: As discussed in the General Response, the revised manuscript now includes rolling-window out-of-sample backtesting, annual regime-based analysis, and exceedance-frequency evaluation to assess predictive usefulness under changing market conditions.
8. The hypothesis “SPY returns are not perfectly normally distributed” is already established and trivially true. A stronger framing is required. Eg. How much does non-normality matter for risk estimation? Which models best capture tail risk? Under what conditions does normality approximation fail? The GMM is evaluated only by Visual fit (histograms, KDE overlays) and In-sample description It does not answer: Does GMM improve VaR accuracy? Does it reduce tail underestimation? Does it generalize out-of-sample? The paper is entirely descriptive:“Here is what the distribution looks like”. But financial modeling requires predictive validity and decision relevance Without prediction, the model is just a smoother curve, not a better model. The paper never demonstrates that its “better-fitting” model improves prediction of any of these quantities (Tail risk (VaR/ES, full return distribution and volatility/extremes). The manuscript evaluates model performance purely in terms of in-sample distributional fit (e.g., visual agreement with histograms and Q-Q plots). However, it is unclear what quantity the model is intended to predict. If the goal is improved risk estimation, the analysis should evaluate out-of-sample predictive performance, such as VaR exceedance rates or likelihood-based scoring under a rolling-window framework. Without such evaluation, the improved fit of the Gaussian Mixture Model remains descriptive rather than practically meaningful.
Response: I agree that simple rejection of normality is not itself a sufficiently strong research contribution. As discussed in the General Response, the revised manuscript was substantially reframed from a descriptive normality-testing study toward practical tail-risk estimation, predictive validation, and model calibration under different market conditions.
The revised manuscript now explicitly investigates how strongly normal distribution assumptions underestimate downside risk, under what conditions model performance deteriorates, and which

alternative models better capture deep-tail behavior across different confidence levels and market regimes.
To address the reviewer's concerns regarding predictive usefulness, the revised paper added rolling-window out-of-sample backtesting, exceedance-frequency analysis, yearly regime-based calibration analysis, AIC/BIC model-selection evaluation, and comparative VaR/ES analysis across Normal, t-distribution, GMM, EVT, and rolling historical models.

9. Please verify that all links to the references point to the correct source.
Response: All DOI links were rechecked and verified to ensure they point to the correct sources.
10. The authors must disclose and acknowledge any assistance received in the preparation of this manuscript, including but not limited to editorial, technical, analytical, or writing support. All such contributions must be clearly stated in the Acknowledgments section.
Response: An "Acknowledgments" section was added to disclose technical guidance, manuscript review support, AI-assisted paper refinement and code debugging support used during research and manuscript preparation.
11. Include enough recent references along with foundational ones. Ensure references directly support your claims. Avoid "padding." Use credible sources (peer-reviewed journals, reputable books, official reports).
Response: All references were reviewed to ensure that they directly support the corresponding claims. Additional recent and relevant sources were added to strengthen the discussion of new contents. All references are from credible sources.

The manuscript is substantially improved. Thank you for addressing my comments. I still have some issues that need to be addressed:
1.Section 3.6 is not doing anything apart form presenting unsurprising information. Please replace section 3.6 with a rolling window sensitivity analysis where you will demonstrate (or not) that your results are not an artifact of the 500 day window. Use additional windows of 250 and 750 days and tabulate the three rolling windows in terms of VaR and ES at 95 and 99%
2.Include a new section 3.10 titled "3.10. Practical Economic Implications of Tail-Risk Underestimation". The content should contain something similar to verbiage that appears at the end of this review. This will translate errors to capital reserve implications, portfolio losses, leverage effects, and regulatory consequences.
3.Convert your statistical risk into actual portfolio performance outcomes. In other words, you already have the SPY csv file. Use that to calculate s and p 500 returns based on each models VaR for the 31 years and compare. This will directly answer the question "Does aggressive Gaussian risk estimation outperform conservative tail-aware strategies over long horizons after accounting for crashes and compounding?" and make it much more interesting because you would be actually predicting returns for 31 years for each of your models. See the attached chatgpt conversation on how to do this.
Suggested section 3.10 verbiage:
The preceding analyses demonstrate that differences among distributional models are not merely statistical but may also have important practical economic consequences for portfolio management, leverage control, and financial risk regulation. In particular, the persistent underestimation of downside risk by the normal distribution at deeper confidence levels suggests that financial institutions relying on thin-tailed assumptions may systematically underestimate potential portfolio losses during periods of market stress.
For example, the rolling-window backtesting analysis showed that the normal-distribution model produced observed 99% VaR exceedances at approximately 2.52 times the theoretically expected frequency, substantially higher than the rolling historical and t-distribution approaches. This indicates that extreme losses occurred far more frequently than predicted under normal assumptions. In practical risk-management settings, such underestimation could lead to insufficient

capital reserves, inadequate margin requirements, or excessive leverage exposure during volatile market environments.

Similarly, the differences in Expected Shortfall (ES) estimates across models imply materially different estimates of potential portfolio drawdowns during crisis periods. At the 99% confidence level, the normal-distribution ES estimate (-3.12%) remained substantially less conservative than the empirical ES (-4.82%), whereas the heavy-tailed models produced estimates much closer to observed downside behavior. For large institutional portfolios, even modest percentage differences in estimated tail losses may correspond to very large differences in required liquidity buffers or potential realized losses during market stress events.

The annual and rolling-window analyses further demonstrate that model performance varies substantially across market regimes. During relatively stable low-volatility periods, the normal distribution often provides acceptable approximation of moderate-tail behavior. However, during stressed environments such as 2008, the normal model increasingly underestimates deeper-tail risk. This regime dependence has important implications for dynamic portfolio allocation and risk monitoring because models calibrated primarily during stable periods may become unreliable precisely when accurate tail-risk estimation becomes most critical.

These findings also have potential regulatory implications. Modern financial regulations, including market-risk capital frameworks, increasingly emphasize stress testing, Expected Shortfall, and backtesting performance because traditional normality-based approaches may fail during extreme market conditions. The present results support the broader view that tail-risk modeling should account for heavy-tailed and time-varying market behavior rather than relying solely on static Gaussian assumptions.

Overall, the results suggest that distributional model selection can materially influence practical estimates of portfolio risk, capital adequacy, leverage exposure, and potential drawdown severity, particularly during periods of elevated volatility and market stress.

Response to comments:

Agreed. Additional discussion was added to the rolling-window analysis, abstract, and conclusions sections to explicitly acknowledge the limitations of static distributional assumptions alone for dynamic tail-risk forecasting.

Thank you very much for your patience and for providing so many constructive comments throughout the review process. Your suggestions substantially improved the overall quality of the research and manuscript.

If you have additional recommendations regarding potential future research directions or methodological improvements, I would greatly appreciate the opportunity to learn from your further feedback.

Thank you for addressing my comments. Your rolling window exceedance ratios of >1.59 means even your best models are underestimating risk. This suggests that static distribution choice alone is insufficient for realistic tail-risk forecasting. Please acknowledge this explicitly in the manuscript. Suggested verbiage "Although the t-distribution and historical simulation substantially improved calibration relative to the normal distribution, all rolling-window models still produced exceedance ratios significantly above the theoretically expected level. This suggests that static distributional assumptions alone are insufficient to fully capture the dynamic and clustered nature of financial tail risk. In particular, volatility clustering, regime transitions, and rapid changes in market conditions may contribute more strongly to tail-risk forecasting error than the choice of unconditional return distribution itself. These findings suggest that conditional volatility models, regime-switching frameworks, or adaptive time-varying approaches may be necessary for more accurate dynamic tail-risk estimation. "

Please also acknowledge in the conclusions section. Suggested line may be "The rolling-window

results further suggest that accurate tail-risk forecasting depends not only on the selection of heavy-tailed distributions, but also on the ability to model time-varying volatility and regime-dependent market behavior.
Also acknowledge in the abstract. Suggested text: “Although heavy-tailed models improved tail-risk calibration relative to the normal distribution, rolling-window analyses showed persistent exceedance underestimation across all models, suggesting that time-varying volatility and regime-dependent behavior play a major role in financial tail-risk dynamics.”

Thank you for addressing my comments. Accepted.