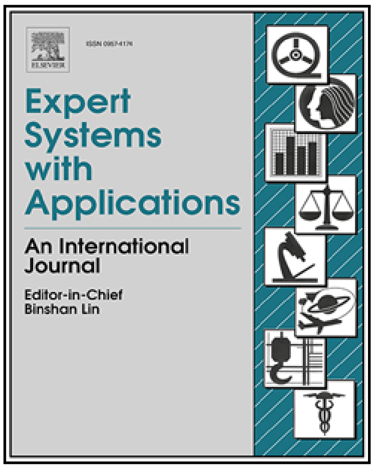
Journal Pre-proof




In the Beginning was the Word: LLM-VaR and LLM-ES

Daniel Traian Pele, Vlad Bolovăneanu, Min-Bin Lin, Rui Ren, Andrei Theodor Ginavar, Bruno Spilak, Alexandru-Victor Andrei, Filip-Mihai Toma, Stefan Lessmann, Wolfgang Karl Härdle

PII: S0957-4174(25)02294-8
DOI: https://doi.org/10.1016/j.eswa.2025.128676
Reference: ESWA 128676

To appear in: *Expert Systems With Applications*

Received date: 16 January 2025
Revised date: 11 June 2025
Accepted date: 17 June 2025

Please cite this article as: Daniel Traian Pele, Vlad Bolovăneanu, Min-Bin Lin, Rui Ren, Andrei Theodor Ginavar, Bruno Spilak, Alexandru-Victor Andrei, Filip-Mihai Toma, Stefan Lessmann, Wolfgang Karl Härdle, In the Beginning was the Word: LLM-VaR and LLM-ES, *Expert Systems With Applications* (2025), doi: https://doi.org/10.1016/j.eswa.2025.128676

**Highlights**

- Propose LLM-VaR and LLM-ES: the first zero-shot, prompt-based estimators for financial Value at Risk and Expected Shortfall using general-purpose LLMs.
- Deliver the first empirical benchmark of GPT-3.5, GPT-4, and GPT-4o for direct tail risk estimation from raw financial time series.
- Pioneer quantile-based financial risk forecasting using general-purpose language models, without retraining or fine-tuning.
- Discover that GPT-3.5 can surpass newer LLMs in tail risk prediction for cryptocurrencies markets.
- Show that LLM-based risk estimators are effective for short-term forecasting, while traditional econometric models remain superior for long horizons.

# In the Beginning was the Word: LLM-VaR and LLM-ES

Daniel Traian Pele[a,b,*], Vlad Bolovăneanu[a], Min-Bin Lin[a], Rui Ren[a,f], Andrei Theodor Ginavar[a], Bruno Spilak[a], Alexandru-Victor Andrei[a], Filip-Mihai Toma[a,f], Stefan Lessmann[a,c], Wolfgang Karl Härdle[a,c,e]

[a] *Bucharest University of Economic Studies, Bucharest, Romania*
[b] *Institute for Economic Forecasting, Romanian Academy, Bucharest, Romania*
[c] *School of Business and Economics, Humboldt University of Berlin, Berlin, Germany*
[d] *California Institute of Technology, Pasadena, CA, USA*
[e] *School of Informatics, University of Edinburgh, Edinburgh, UK*
[f] *University of Augsburg, Augsburg, Germany*



**Abstract**

This study introduces **LLM-VaR** and **LLM-ES**, novel risk estimation metrics that utilize general-purpose large language models (LLMs) for the forecasting tasks of Value at Risk (VaR) and Expected Shortfall (ES) in a zero-shot setting. Building on the input encoding mechanism of the LLM-Time framework, we extend its application by defining new financial risk measures and performing an empirical evaluation of three generations of GPT models, GPT-3.5, GPT-4 and GPT-4o, versus advanced benchmark models such as GARCH with Student innovations and EWMA with Dynamic Conditional Score (DCS).

Financial time series are encoded as numerical strings, allowing for model-free inference without requiring retraining. Results show that LLMs perform well when short rolling windows are used, particularly in volatile markets like cryptocurrencies. GPT-3.5 frequently outperforms or matches the performance of newer models, raising questions about model complexity, align-

*Corresponding author

*Email addresses:* danpele@ase.ro (Daniel Traian Pele), vlad.bolovaneanu@ase.ro (Vlad Bolovăneanu), linminbin0593@gmail.com (Min-Bin Lin), rui.ren.magic@gmail.com (Rui Ren), ginavarandrei19@stud.ase.ro (Andrei Theodor Ginavar), bruno.spilak@gmail.com (Bruno Spilak), andrei1victor23@stud.ase.ro (Alexandru-Victor Andrei), mihai.toma@fabiz.ase.ro (Filip-Mihai Toma), stefan.lessmann@hu-berlin.de (Stefan Lessmann), haerdle@hu-berlin.de (Wolfgang Karl Härdle)

ment, and biases. In contrast, performance deteriorates with longer windows, where the econometric models prove more reliable. Our findings demonstrate the potential of general-purpose LLMs as adaptive tools for short-horizon financial risk assessment and contribute a first-of-its-kind benchmark for LLM-based VaR/ES estimation.

*Keywords:* Value at Risk, Expected Shortfall, GPT, LLM-VaR, LLM-ES, Large Language Models

## 1. Introduction

Value at Risk (VaR) and Expected Shortfall (ES) are cornerstone metrics in financial risk management, offering quantitative estimates of potential portfolio losses under adverse market conditions. Although VaR provides a threshold-based loss estimate at a given confidence level, ES captures the average loss beyond that threshold, delivering a more comprehensive view of tail risk. Traditional methods for estimating VaR and ES—such as parametric models, GARCH frameworks, historical simulations, or Monte Carlo simulations—often suffer from rigid assumptions, limited adaptability, and high computational demands, particularly in dynamic market environments.

Recent advances in artificial intelligence, particularly in large language models (LLMs) based on Transformer architectures, have significantly expanded our ability to model sequential data. General-purpose LLMs, such as GPT-3.5 and GPT-4, have demonstrated strong performance across domains, from forecasting and anomaly detection to decision support. Their zero-shot and few-shot capabilities allow them to generalize to new tasks with minimal supervision, making them attractive for real-time financial applications where adaptability is crucial (OpenAI, 2023).

Despite their growing adoption, the use of general-purpose LLMs for structured, numerically grounded tasks, such as financial risk estimation, remains underexplored. Most existing studies focus on sentiment analysis, language understanding, or feature extraction from unstructured data. In contrast, traditional models still dominate VaR and ES estimation, even though they struggle with nonstationarity and nonlinearities without extensive recalibration.

This raises a timely question: *Can general-purpose LLMs be reliably adapted for real-time estimation of financial tail risk, and under what conditions could they outperform or complement traditional models?*

Addressing this question is important for both researchers and practitioners. If LLMs can generate robust VaR and ES estimates directly from price data, they could serve as scalable, model-free tools that require little calibration. This would improve response to regime changes and reduce dependence on market-specific tuning. Additionally, understanding their strengths and limitations can inform the design of hybrid architectures that integrate the statistical reliability of traditional models with the flexibility of neural approaches.

This study investigates the feasibility of using general-purpose LLMs for real-time financial risk estimation, introducing two novel risk metrics: **LLM-VaR** and **LLM-ES**. These are derived from LLM outputs applied to encoded financial time series and are tested across three generations of OpenAI's GPT models (3.5, 4, and 4o). We build on the LLMTime framework (Gruver et al., 2024) for time-series encoding and model interaction and extend this framework for financial risk assessment through tailored prompting, parameter sensitivity analysis (for the temperature parameter), and robust benchmarking against both standard and extended GARCH(1,1) and EWMA models.

Our empirical evaluation focuses on short-horizon, high-volatility settings—such as cryptocurrencies—and assesses trade-offs in cost, scalability, and data governance.

As financial supervisory authorities explore AI-driven solutions (see, for example, European Central Bank, McCaul (2024)), an LLM-based VaR/ES tool that operates without retraining could greatly improve the flexibility and timeliness of risk reporting and early-warning mechanisms.

The structure of the paper is as follows: Section 2 reviews related work on financial risk estimation, time-series forecasting, and LLM applications in finance; Section 3 details our methodology, including the LLM-VaR and LLM-ES concepts and testing framework; Section 4 presents the datasets and empirical results; Section 5 discusses limitations and practical considerations; Section 6 provides a detailed discussion of model behavior and broader implications; and Section 7 concludes.

Data and replication code are accessible via Quantlet.com. A courselet on this topic is available at Quantinar.com.

## 2. Related work

The emergence of Large Language Models (LLMs) has the potential to significantly reshape how financial institutions assess risk, particularly with respect to Value at Risk (VaR) and Expected Shortfall (ES). The ability of LLMs to process and analyze large volumes of structured and unstructured data enables more tailored and dynamic models for financial risk management.

Previous research has provided a diverse range of models for estimating VaR and ES, both of which are central to modern financial risk management. Traditional methods include historical simulation, parametric models, Monte Carlo simulation, and the GARCH family of models. These approaches are grounded in strong theoretical underpinnings but often rely on assumptions, such as normality, stationarity, or specific volatility structures, that may not hold in volatile or rapidly shifting market regimes. Moreover, they typically require substantial computational effort for calibration and do not generalize well to new asset classes or structural breaks.

In response to these limitations, a growing body of work has explored the use of machine learning to improve tail-risk estimation. For example, Qiu et al. (2024) proposed stateful recurrent neural networks that outperform conventional models in one-day VaR and ES prediction, while Wang et al. (2024a) introduced a hybrid deep learning framework combining quantile regression with Mogrifier RNNs and GANs to better simulate and forecast extreme losses. Further, Fatouros et al. (2023) introduced the DeepVaR architecture, a probabilistic deep neural network that improves estimation accuracy for high quantiles of return distributions.

Transformer-based models have further advanced the field of time series modeling. Architectures such as Informer, Autoformer, and Fedformer use attention mechanisms to effectively capture long-term dependencies (Zhou et al., 2022). These serve as the foundation for models like TimesFM and Salesforce's Moirai, which are pretrained on billions of time steps and applied to financial forecasting in a zero-shot or fine-tuned manner (Nie et al., 2024). Foundation models have recently been applied to financial time series forecasting tasks, demonstrating strong performance in volatility and tail risk estimation. For example, Goel et al. (2025a) introduced a time-series foundation model for VaR forecasting, comparing Google's TimesFM model—both in zero-shot and fine-tuned variants—to traditional methods such as GARCH and Generalized Autoregressive Score (GAS). Using 19 years of S&P 100 returns and over 8.5 years of out-of-sample backtesting, they found that fine-tuning significantly improved performance across multiple quantiles (0.01 to 0.1), often outperforming conventional econometric models in actual-over-expected ratios and quantile score loss. In a related study, Goel et al. (2025b) demonstrated that the same model architecture, when fine-tuned for realized volatility, also exceeded the forecasting accuracy of classical volatility models. These findings underscore the adaptability of foundation models for both central and tail-risk forecasting, though they still require task-specific tuning to perform optimally.

The recent literature further suggests that LLMs can enhance the statistical approaches traditionally employed in financial analysis. For instance, Trachova and Lysak (2025) emphasize the role of LLMs in combining narrative and quantitative data, allowing for improved risk and fraud detection in financial reporting. This integration is crucial for deriving accurate estimates of VaR, as traditional methods often rely heavily on historical data alone, which can lead to underestimations during volatile market conditions. Similarly, Li et al. (2025) highlight that LLMs can navigate financial documents to uncover insights that directly impact risk assessments, helping firms to calculate VaR more effectively while understanding the driving factors behind these risks. LLMs can also analyze sentiment and contextual information from financial news and reports, further informing risk evaluations and enabling deeper integration with existing financial systems (Li et al., 2025).

Advancements in LLM technology have extended their application to market forecasting and risk assessment. Liu (2025) identifies how Financial Language Models (FinLLMs) are applied to sentiment analysis and risk assessments, underscoring their utility in recognizing market patterns that inform VaR and ES calculations. By leveraging deep learning techniques and domain-specific fine-tuning, practitioners can create models that adapt to evolving market conditions and derive more accurate risk measures. In practical applications, LLMs have been shown to outperform traditional models in market analysis tasks, enabling financial analysts to develop more resilient risk mitigation strategies and explore dynamic risk limits beyond the static models typically used for VaR and ES (Lee, 2025; Lagasio et al., 2025).

Recent work has also introduced hybrid and multimodal risk modeling pipelines that combine structured financial data with unstructured sources such as audio and text. For example, RiskLabs (Cao et al., 2025) proposes a comprehensive framework that leverages large language models to predict financial risk by fusing data from earnings conference calls, time series, and contextual news. In a parallel development, FinTral (Bhatia et al., 2024) introduces a suite of multimodal LLMs built on the Mistral-7B backbone, supporting reasoning over textual, tabular, numerical, and image data simultaneously.

Frameworks such as LLMTime (Gruver et al., 2024) have demonstrated how general-purpose LLMs can be prompted with tokenized time series data for forecasting tasks, although most applications so far rely on either (i) fine-tuning on financial data, or (ii) augmenting models with domain-specific inputs.

Despite the rapid progress, little is known about the capability of general-purpose LLMs to forecast financial risk metrics such as VaR and ES in a zero-shot setting, using only structured historical data encoded as language. To our knowledge, no prior work evaluates LLMs' raw ability to forecast risk measures without retraining or additional financial supervision. Our study fills this gap by benchmarking LLM-generated VaR and ES forecasts - produced via prompt interaction only - against standard and extended GARCH and EWMA baselines, across multiple model generations and market conditions.

## 3. Methodology

In this study, we employ three generations of general-purpose GPT models (3.5, 4, and 4o) [1] within the LLMTime framework to estimate VaR and ES through a zero-shot forecasting approach, which requires no task-specific retraining. This approach builds on the broad pre-training of GPT models to facilitate adaptability and responsiveness to real-time market conditions, making it well-suited for dynamic financial environments.

Our methodology encodes financial asset log-returns as sequential inputs for the LLMs, leveraging the models' extensive cross-domain knowledge. Here, we adapt LLMs for financial risk assessment by encoding numerical returns as string tokens, enabling the model to process financial data in a manner similar to natural language. Each LLM generates a probability distribution for future returns informed by historical data. From this distribution, we derive VaR as the $\alpha$-quantile and ES as the conditional expectation of log-returns, given an exceedance beyond the VaR threshold. This process supports the models' flexibility in risk prediction, providing scalability and eliminating the need for recalibration, thereby enabling real-time updates in risk estimation.

[1] In OpenAI terminology: gpt-3.5-turbo-instruct, gpt-4-turbo, gpt-4o, see https://platform.openai.com/docs/models.

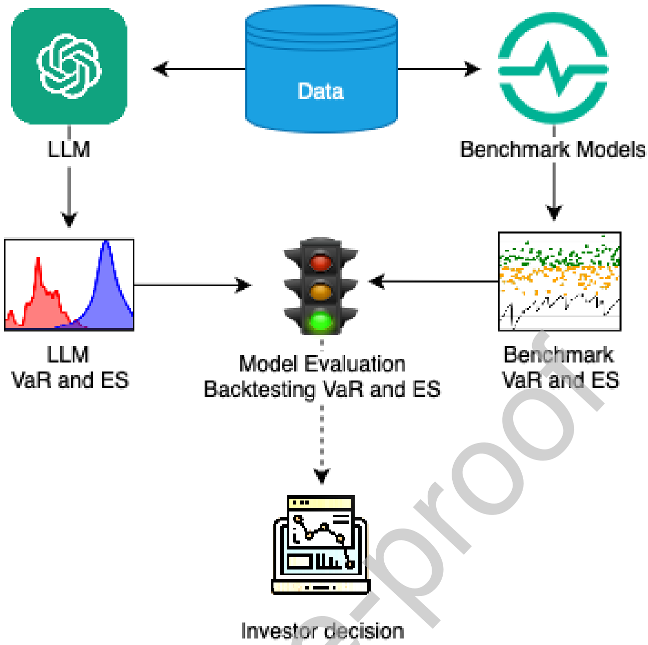



Figure 1: LLM VaR and ES prediction and evaluation system

To evaluate the effectiveness of these LLM-based risk measures, we apply established backtesting procedures, including the Kupiec Proportion of Failures (POF) test, the Traffic Light test, and Christoffersen's Conditional Coverage test for VaR. For ES backtesting, we used the $Z_2$ and $Z_3$ tests from Acerbi and Székely (2014). These allow comprehensive validation of both exceedance frequency and independence. Traditional models, such as GARCH and historical simulation, serve as benchmarks, allowing us to assess whether general-purpose LLMs within the LLMTime framework can produce comparable or superior risk estimates in terms of accuracy, reliability, and computational efficiency.

Figure 1 illustrates the system architecture for LLM-based VaR and ES prediction and evaluation.

*3.1. LLM Architecture*

Given a sequence of financial returns $\{r_1, r_2, \ldots, r_n\}$, each return is transformed into a string and preprocessed according to LLMTime (e.g., "0.598" $\rightarrow$ "59", "-0.209" $\rightarrow$ "-21") to be compatible with natural language processing architectures. Each stringified return is then decomposed into *tokens*[2]—the smallest unit of data that the language model processes—and mapped to a continuous vector space via an embedding matrix $E$, producing embeddings $\mathbf{e}_i$ for each token at time $i$. The LLM processes these embeddings while maintaining temporal structure using positional encodings, leading to initial token representations $\mathbf{z}_i^0 = \mathbf{e}_i + \mathbf{p}_i$[3] (Ahmed et al., 2023).

Dependencies across time points are captured using attention mechanisms within Transformer blocks, each consisting of a multi-head self-attention layer and a feed-forward neural network. In each Transformer block, the self-attention mechanism enables the model to assign weights to different tokens by calculating attention scores for each pair of tokens. Specifically, each token in the sequence is associated with query, key, and value vectors. The attention weights for tokens $i$ and $j$ are calculated as:

$$\alpha_{ij} = \exp\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_k}}\right) \Bigg/ \sum_{j'=1}^{n} \exp\left(\frac{\mathbf{q}_i^\top \mathbf{k}_{j'}}{\sqrt{d_k}}\right), \tag{1}$$

where $d_k$ is the dimension of the key vector. The output at each position $i$ is derived by weighting the value vectors from all positions in the sequence: $\mathbf{h}_i = \sum_{j=1}^{n} \alpha_{ij} \mathbf{v}_j$. After passing through multiple Transformer blocks, the model generates final hidden representations, which are then used to predict the next token.

The LLM models the conditional distribution of the next token given past observations as $\mathrm{P}(r_t \mid r_{t-n}, \ldots, r_{t-1})$, enabling the estimation of risk measures.

This approach supports flexible, adaptive financial risk modeling, where the LLM predicts the expected return $\hat{r}_t$ based on the conditional expectation given past tokens (representing past returns): $\hat{r}_t = \mathsf{E}[r_t \mid r_{t-n}, \ldots, r_{t-1}]$.

For chat models (GPT-4 and GPT-4o), we have adopted the prompt suggested by Gruver et al. (2024). We have chosen the `Instruct` version of GPT-3.5, also employed by the LLMTime authors on their Github page (Gruver, 2025).

[2]Usually one token per number, as GPT-3.5 and onwards include separate tokens for all numbers from 0 to 999.

[3]GPT-4 and GPT-4o do not disclose their architectural choices, therefore, the implementation of positional embeddings may differ.

*3.2. LLM-Based Risk Measures*

Consider a financial asset or portfolio with log-returns denoted by $r_t = \log P_t - \log P_{t-1}$ at time $t$, where $P_t$ represents the closing price of the asset. The VaR ($\mathrm{VaR}_t^\alpha$) at confidence level $\alpha$, for a one-period horizon, conditional on information available at $t-1$, is defined as the maximum expected loss not exceeded with probability $\alpha$. Formally:

$$\mathrm{P}(r_t \leq \mathrm{VaR}_t^\alpha) = \alpha \quad \Leftrightarrow \quad \mathrm{VaR}_t^\alpha = -\mathcal{F}_{t-1}^{-1}(\alpha), \tag{2}$$

where $\mathcal{F}_{t-1}$ represents the cumulative distribution function of log-returns, conditional on information at $t-1$.

ES ($\mathrm{ES}_t^\alpha$), or Conditional Value at Risk (CVaR), quantifies the average loss conditional on returns falling below the $\mathrm{VaR}_t^\alpha$ threshold, thereby capturing tail risk beyond the VaR limit. It is expressed as:

$$\mathrm{ES}_t^\alpha = -\mathsf{E}[r_t \mid r_t \leq \mathrm{VaR}_t^\alpha] = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_t^\gamma \, d\gamma. \tag{3}$$

*3.2.1. LLM-VaR and LLM-ES*

Within the context of LLM forecasting, we define LLM-VaR and LLM-ES for a given model $M$ as follows: $\mathrm{VaR}_t^{\alpha;M}$ and $\mathrm{ES}_t^{\alpha;M}$ represent the model's estimate of VaR and ES, respectively:

$$\mathrm{VaR}_t^{\alpha;M} = -\hat{\mathcal{F}}_t^{M;-1}(\alpha), \quad \mathrm{ES}_t^{\alpha;M} = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_t^{\gamma;M} \, d\gamma, \tag{4}$$

where $\mathcal{Y}_t^M = \{\hat{r}_t^{i;M}\}_{i=1}^n$ represents the set of forecast returns generated by LLM $M$, and $\hat{\mathcal{F}}_t^M$ denotes the empirical cumulative distribution function derived from these forecast returns, $\mathcal{Y}_t^M$.

Forecast returns are generated as $\mathcal{Y}_t^M = f^M(\mathcal{X}_{t-1}; \Theta^M)$, where $\mathcal{X}_{t-1}$ includes relevant input features (e.g., historical returns) up to time $t-1$, $\Theta^M$ signifies the model parameters, and $f^M$ is the predictive function of the LLM. As noted in Gruver et al. (2024), LLMs adapt to time series applications by encoding numerical data as sequences of strings, thus leveraging the model's linguistic architecture for structured, predictive outputs.

This approach offers a novel methodology for risk estimation, as the flexibility inherent in LLMs enables adaptation to various financial contexts without requiring task-specific retraining, thereby providing a scalable solution for dynamic risk management applications.

*3.2.2. Estimation Algorithm*

We utilize GPT-3.5 Turbo, GPT-4, and GPT-4o to estimate VaR and ES. This is accomplished through a rolling window approach with window length $w$, using historical log-returns as inputs. At each time step $t$, we apply a rolling window of past returns $\mathcal{X}_{t-1} = \{r_{t-1}, r_{t-2}, \dots, r_{t-w}\}$ as input for the LLM $M$. The model generates a series of samples representing potential realizations of the next log-return $\hat{r}_t^M$, which are used to construct the empirical cumulative distribution function $\hat{\mathcal{F}}_t^M$. VaR and ES estimates for time $t$ are derived from this empirical distribution.

Each model is configured with hyperparameters $\Theta^M = \{\tau = 0.7, \alpha_{LLM} = 0.95, \beta_{LLM} = 0.35, \pi = 2\}$, as outlined by Gruver et al. (2024). These hyperparameters are critical in fine-tuning the models for optimal performance:

- $\tau$ (temperature) controls the randomness in model outputs; higher values increase variability in predictions, which can enhance exploration. We run an extensive analysis to assess its impact in Section 4.3.
- $\alpha_{LLM}$ and $\beta_{LLM}$ calibrate the model's sensitivity to numerical inputs, ensuring it effectively manages both large and small values.
- $\pi$ determines the granularity of tokenizing numerical data, refining the precision with which log-returns are encoded. Somewhat counterintuitively, a small value (2 or 3) is preferred because of the trade-off between numerical precision (which can induce noise) and general signal characteristics; see Bianchi et al. (2025) for a possible explanation. This trend is further reflected in recent advances in large language model training, where reduced numerical precision—such as 4-bit or 8-bit quantization—has been successfully employed to balance efficiency and signal integrity (see, e.g., (DeepSeek-AI et al., 2025)).

Gruver et al. (2024) recommends adding a space between the return digits for GPT-3, which uses a different tokenizer than GPT-3.5, 4, and 4o (OpenAI, 2025c). We noticed in our experiments that "gluing" the digits of each return, that is, setting `bit_sep = "`, is a suitable approach.

This set of hyperparameters is selected to balance predictive accuracy and computational efficiency, thus optimizing the estimation of VaR and ES within the LLM framework.

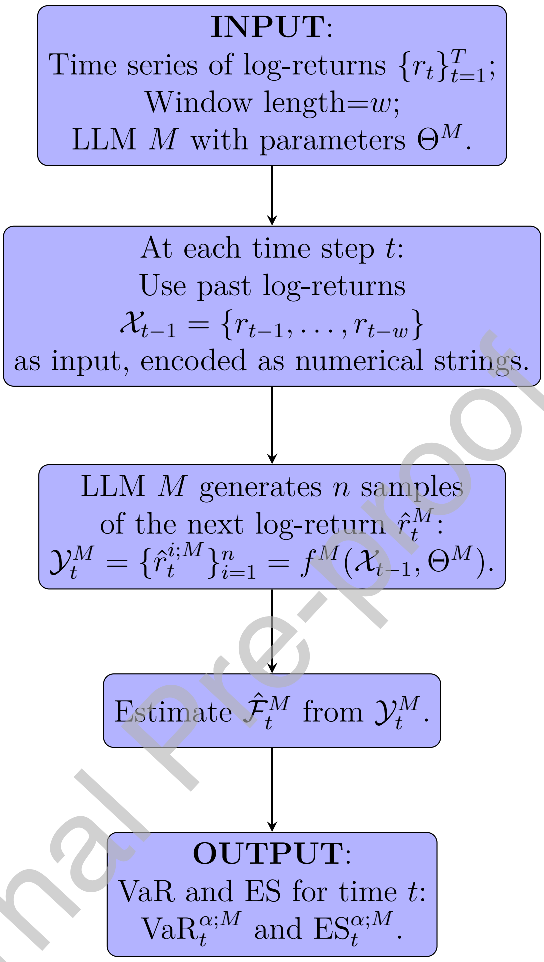



Figure 2: Rolling window approach for LLM-based risk measures.

### *3.3. Benchmark Models for Risk Estimation*

To rigorously evaluate the effectiveness of our proposed LLM-based methods for VaR and Expected ES, we benchmark them against two advanced versions of widely used models in financial risk estimation: the GARCH(1,1) model and the Exponentially Weighted Moving Average (EWMA) model. Our aim is to make the benchmarks as robust as possible, thereby subjecting the LLM-based methods to a stringent test. To this end, we extend beyond standard, or "vanilla" implementations of GARCH and EWMA, incorporating enhancements that adapt these models to handle complex market dynamics more effectively. The GARCH(1,1) model (Bollerslev, 1986) effectively captures volatility clustering in financial returns, with high-volatility periods tending to follow one another. The model updates volatility dynamically based on past returns and volatility. We further enhance GARCH by using a Local Parametric Approach (LPA) (Spokoiny, 1998) for detecting and adjusting to structural market shifts (Spilak and Härdle, 2022).

For the EWMA model, we incorporate the Dynamic Conditional Score (DCS) framework (Creal et al., 2013), also called the Generalized Autoregressive Score (GAS) framework. This enhanced EWMA model dynamically adjusts volatility based on market conditions and accounts for heavy tails under a Student's t-distribution. The DCS framework is similarly applied to GARCH, allowing it to better capture sudden changes in market dynamics. Additional details on these benchmarks are provided in Appendix A.

These robust benchmark models are designed to offer a high-performance baseline against which we compare the LLM-based approaches. By fortifying these traditional models, we aim to create the most challenging possible conditions, testing whether LLMs can provide additional flexibility and adaptability in capturing dynamic risk factors in financial markets.

For the benchmark models, we compute VaR and ES based on the conditional distribution of returns as follows (we used the formulation in McNeil et al. (2005) for ES):

$$\mathrm{VaR}_t^{\alpha} = -\widehat{\sigma}_t q_{\alpha}, \tag{5}$$

and

$$\mathrm{ES}_t^{\alpha} = \begin{cases} \frac{\widehat{\sigma}_t}{\alpha}\phi(\Phi^{-1}(1-\alpha)), & Z_t \sim \mathcal{N} \\ \frac{\widehat{\sigma}_t}{\alpha}\left(g_{\nu}(t_{\nu}^{-1}(1-\alpha))\right)\frac{\nu+(t_{\nu}^{-1}(1-\alpha))^2}{\nu-1}, & Z_t \sim t_{\nu} \end{cases}, \tag{6}$$

where $Z_t$ is defined in Appendix A, $\phi, \Phi^{-1}$ are the probability density function (PDF) and inverse cumulative distribution function (CDF) of the standard Normal distribution, $t_{\nu}, g_{\nu}$ the PDF and inverse CDF of a standard Student distribution with $\nu$ degrees of freedom, and $\widehat{\sigma}_t$ the predicted volatility.

*3.4. Backtesting VaR*

Backtesting is a crucial tool for assessing the accuracy and robustness of VaR models in risk management. This section outlines three widely recognized backtesting methods: the Kupiec Test, the Traffic Light Approach, and Christoffersen's Conditional Coverage (CC) test, each providing distinct insights into model performance and reliability.

**The Kupiec Test** (Kupiec, 1995), also known as the **Proportion of Failures (POF) test**, examines whether the observed frequency of VaR breaches (exceedances) is consistent with the expected probability of exceedance, $\alpha$, specified by the model's confidence level. Let $p$ denote the true probability of a VaR exceedance in the population. The null hypothesis, $H_0 : p = \alpha$, is tested using the POF likelihood ratio statistic:

$$LR_{\mathrm{POF}} = -2\log\left\{\frac{(1-\alpha)^{N-x}\alpha^x}{\left(1-\frac{x}{N}\right)^{N-x}\left(\frac{x}{N}\right)^x}\right\}, \tag{7}$$

where $x$ represents the number of observed exceedances, $N$ is the total sample size, and $\alpha$ is the model's confidence level. If $LR_{\mathrm{POF}}$ exceeds the critical value, the VaR model may be deemed inadequately calibrated. For scenarios with zero exceedances ($x = 0$), the POF test simplifies to:

$$LR_{\mathrm{POF}} = -2\log\left\{(1-\alpha)^N\right\}, \tag{8}$$

enabling an assessment of whether a lack of failures aligns with the expected exceedance rate, $\alpha \cdot N$. A well-calibrated model should exhibit an exceedance rate close to the targeted confidence level.

**The Traffic Light Approach** (Basel Committee on Banking Supervision, 1996) classifies VaR performance into three distinct zones: Green (acceptable performance), Yellow (potential issues), and Red (unacceptable). The exceedance indicator $X_t^{\mathrm{VaR}}(\alpha)$ for a one-period-ahead VaR estimate is defined as:

$$X_t^{\mathrm{VaR}}(\alpha) = \mathbf{1}_{\{r_t \leq -\,\mathrm{VaR}_t^\alpha\}}, \tag{9}$$

where $\mathbb{1}$ denotes the indicator function. For a sufficiently large sample size $N$, the cumulative exceedance count $X_N^{\mathrm{VaR}}(\alpha)$ approximates a normal distribution:

$$X_N^{\mathrm{VaR}}(\alpha) \sim \mathcal{N}(N\alpha, N\alpha(1-\alpha)). \tag{10}$$

The standard normal transform $z$ is used to determine the traffic light zone: Green if $\Phi(z) < 0.95$, Yellow if $0.95 \leq \Phi(z) < 0.9999$, and Red if $\Phi(z) \geq 0.9999$, where $\Phi$ is the cumulative distribution function (CDF) of the standard Normal distribution(Alexander and Dakos, 2023).

**Christoffersen's Conditional Coverage (CC) Test** (Christoffersen, 1998) extends the POF test by examining both the frequency and independence of VaR exceedances. The test evaluates whether exceedances are independently distributed over time, thus capturing any clustering of failures. The likelihood ratio for the independence test is given by:

$$LR_{\text{CCI}} = -2\log\left(\frac{(1-\pi)^{n_{00}+n_{10}}\pi^{n_{01}+n_{11}}}{(1-\pi_0)^{n_{00}}\pi_0^{n_{01}}(1-\pi_1)^{n_{10}}\pi_1^{n_{11}}}\right), \tag{11}$$

where $n_{ij}$ represents the count of transitions between periods of failure and non-failure (e.g. $i = 1, j = 0$ represents the transition from a failure to a non-failure state), with $\pi_0$, $\pi_1$, and $\pi$ denoting the transition probabilities. The combined Conditional Coverage test statistic is defined as:

$$LR_{\text{CC}} = LR_{\text{POF}} + LR_{\text{CCI}}, \quad LR_{\text{CC}} \sim \chi^2(2). \tag{12}$$

For cases where zero exceedances are observed ($x = 0$), the CC test reduces to the POF test, setting $LR_{\text{CCI}} = 0$.

*3.5. Backtesting ES*

We apply two robust ES backtesting methods from Acerbi and Székely (2014)[4].

**$\mathbf{Z_2}$ Test for ES**, widely applied in the studies such as Lazar and Zhang (2019) and Clift et al. (2016), evaluates both the frequency and severity of ES breaches. This dual evaluation helps capture scenarios where VaR may not sufficiently reflect the extreme losses in the distribution tail. The $Z_2$ statistic, formulated based on the unconditional ES definition, is calculated as:

$$Z_2 = \sum_{t=1}^{T} \frac{I_t r_t}{T\alpha\, \text{ES}_t^{\alpha}} + 1, \tag{13}$$

where $I_t = \mathbb{1}_{\{r_t \leq -\text{VaR}_t^{\beta}\}}$, with $\alpha = 0.025$ and $\beta = 0.01$ for testing 2.5% ES. Under the null hypothesis $H_0$, which assumes unbiased ES estimates, $Z_2$ has

[4]We explored tests requiring correctly specified VaR for the null hypothesis, including the generalized traffic light approach by Costanzino and Curran (2018) and the comprehensive coverage test by Costanzino and Curran (2015). However, uncalibrated VaR (as observed in certain GPT-4 experiments) yielded misleadingly favorable p-values. Therefore, our selection emphasizes tests that enhance result robustness without this assumption.

an expected value of zero, formally:

$$\begin{aligned} H_0: &\quad P_t^{[\alpha]} = F_t^{[\alpha]}, \quad \forall t, \\ H_1: &\quad \mathrm{ES}_t^{\alpha,F} \geq \mathrm{ES}_t^{\alpha} \text{ for all } t \text{ and strictly greater for some } t, \\ &\quad \mathrm{VaR}_t^{\beta,F} \geq \mathrm{VaR}_t^{\beta} \text{ for all } t. \end{aligned} \tag{14}$$

In (14), $P_t$ represents the estimated conditional distribution, while $F_t$ is the true conditional distribution. The function $P_t^{[\alpha]} = \min(1, P_t(x)/\alpha)$ denotes the tail of $P_t$, populated only by exceedances. Suffix $F$ denotes the true values derived from $F_t$.

Deviations of $Z_2$ below zero, particularly with $Z_2 < Z_2^* = -0.7$ (the 5% critical threshold, stable across tests), indicate consistent overestimation of tail risk, prompting rejection of $H_0$. Following Clift et al. (2016), we use a simulation of size $M = 20,000$ to obtain p-values.

$\mathbf{Z_3}$ **Test for ES** complements the analysis, focusing on the ranks $U_t = F_t(r_t)$, which ideally follow an i.i.d. $\mathcal{U}(0,1)$ distribution. The vector $U = \{U_t\}$ is used to re-estimate ES across previous days, and the average is compared with an i.i.d. uniform average:

$$Z_3 = -\frac{1}{T}\sum_{t=1}^{T} \frac{\widehat{\mathrm{ES}}_\alpha^{(T)}(P_t^{-1}(U))}{\mathsf{E}_V\left[\widehat{\mathrm{ES}}_\alpha^{(T)}(P_t^{-1}(V))\right]} + 1, \tag{15}$$

where $V$ is a vector of $T$ i.i.d. $\mathcal{U}(0,1)$, and $\widehat{\mathrm{ES}}_\alpha^{(T)}$ denotes the empirical[5] ES, based on a vector of $N$ i.i.d draws $\vec{Y} = \{Y_i\}$:

$$\widehat{\mathrm{ES}}_\alpha^{(T)}(Y) = -\frac{1}{\lfloor T\alpha \rfloor}\sum_{i=1}^{\lfloor T\alpha \rfloor} Y_{i:T}. \tag{16}$$

The denominator is approximated as:

$$-\frac{T}{\lfloor T\alpha \rfloor}\int_0^1 I_{1-p}(T - \lfloor T\alpha \rfloor, \lfloor T\alpha \rfloor) P_t^{-1}(p)\, dp, \tag{17}$$

where $I_{1-x}(a,b)$ is the regularized incomplete Beta function. We employ

[5] $\lfloor x \rfloor$ is the integer part of $x$ and $Y_{i:N}$ denotes order statistics.

Simpson's rule with 1,000 intervals to approximate this integral.

Each day's contribution ideally equals 1; thus, for $Z_3$, we expect $\mathsf{E}_{H_0}[Z_3] = 0$ and $\mathsf{E}_{H_1}[Z_3] < 0$.

Because this test does not rely on estimated VaR and ES, its assumptions pertain to the full distribution:

$$\begin{aligned} H_0: &\quad P_t = F_t, \quad \forall t, \\ H_1: &\quad P_t \succeq F_t, \quad \text{for all } t \text{ and strictly } P_t \succ F_t \text{ for some } t. \end{aligned} \tag{18}$$

As the $Z_3$ test is computationally intensive, we conducted fewer simulations, yielding consistent results. This study utilizes 1,000 simulations. For LLMs, we apply Kernel Density Estimation with a Gaussian kernel and automatic bandwidth selection on each day's predicted log-return distribution.

## 4. Data and Empirical Results

### *4.1. Data*

This study analyzes data spanning from October 1, 2021, to March 13, 2024, a period specifically chosen to ensure the data falls outside the GPT-3.5 Turbo model's training cutoff (September 2021, OpenAI (2024)). By starting the dataset on October 1, 2021, we ensure that the analysis incorporates unseen data not part of the LLM's pre-trained knowledge.

Our dataset includes daily log-returns for nine different indices, covering diverse fields such as cryptocurrencies, stocks, clean energy, bonds, and commodities (see Table 1). The CRIX index, representing cryptocurrencies, exhibits a significantly higher number of daily log-returns compared to other assets, due to the continuous 24/7 nature of cryptocurrency trading, unlike traditional markets that observe fixed trading hours and holidays.

Table 1: Assets used for analysis.

| Nr. | Symbol | Name | Source | # daily log-returns |
|---|---|---|---|---|
| 1 | CRIX | Cryptocurrency Index | Royalton | 895 |
| 2 | S&P 500 | Standard and Poor's 500 | Refinitiv | 614 |
| 3 | SPGTCLTR | S&P Global Clean Energy Index | Refinitiv | 638 |
| 4 | STOXX | STOXX Europe 600 Index | Refinitiv | 630 |
| 5 | CACT | CAC All-Tradable | Refinitiv | 629 |
| 6 | GDAXI | Deutsche Boerse DAX Index | Refinitiv | 627 |
| 7 | CBU0.L | iShares $ Treasury Bd 7-10y ETF USD | Refinitiv | 616 |
| 8 | FTSE100 | Financial Times Stock Exchange 100 Index | Refinitiv | 616 |
| 9 | DJCI | Dow Jones Commodity Index | Refinitiv | 614 |

For each asset, the LLM-based forecasting approach described in Section 3.2.2 was implemented using a rolling-window methodology. The window length $w$ varied from 30 to 150 days, specifically $w \in \{30, 45, 60, 90, 120, 150\}$. At each time point $t$, we simulated $n = 2^{10} = 1024$ values for the next day's log-return, represented as $\mathcal{Y}_t^M = \{\hat{r}_t^{i;M}\}_{i=1}^n = f^M(\mathcal{X}_{t-1}, \Theta^M)$, based on past log-returns $\mathcal{X}_{t-1} = \{r_{t-1}, r_{t-2}, \ldots, r_{t-w}\}$ encoded as numerical strings.

The choice of $n = 1024$ simulations is based on two key factors: the API limit on completions (128 per request, as set by OpenAI) and the need for a substantial sample size to estimate the empirical cumulative distribution function (ECDF) of forecast log-returns. To achieve this, we generated $128 \times 8$ values per time step, providing sufficient data for robust statistical analysis.

For benchmarks, we evaluate GARCH models under both normal and Student's t-distributions across 120- and 250-day horizons, specifically including GARCH Normal, GARCH DCS Normal, and GARCH DCS Student variants, as well as GARCH LPA without a fixed time horizon. EWMA benchmarks are conducted using normal and Student's t-distributions across 80- and 120-day horizons, specifically EWMA Normal, EWMA DCS Normal, and EWMA DCS Student.

All experiments have been performed on a server with 2 Intel(R) Xeon(R) Gold 6342 2.80GHz CPUs and 256 GM RAM.

### 4.2. Backtesting Results

In this section, we present the backtesting results for the 1% VaR and the 2.5% ES, as recommended by the Basel Committee on Banking Supervision (BCBS) (Basel Committee on Banking Supervision, 1996). The main results are presented here, while the remaining tables and charts can be found in Appendix B and Appendix C.

#### 4.2.1. VaR Backtesting Results

This section evaluates the performance of LLM models compared to classical approaches such as GARCH and EWMA in predicting 1% VaR across multiple assets. The aim is to assess whether LLM models offer competitive performance in capturing risk across diverse markets.

Table 2: Percentage of failures (POF) for 1% VaR (%).

| Model | CACT | DJCI | FTSE | CRIX | CBU | SP500 | STOXX | SPGTCLTR | GDAXI |
|---|---|---|---|---|---|---|---|---|---|
| GPT-3.5.30 | 1.015 | 1.042 | 1.536 | 0.935 | 1.024 | 0.871 | 1.014 | 0.500 | 1.019 |
| GPT-3.5.45 | 0.521 | 0.357 | 0.525 | 0.713 | 0.175 | 0.000 | 0.520 | 0.342 | 0.523 |
| GPT-3.5.60 | 0.178 | 0.183 | 0.540 | 0.363 | 0.180 | 0.000 | 0.000 | 0.351 | 0.000 |
| GPT-3.5.90 | 0.000 | 0.000 | 0.380 | 0.126 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-3.5.120 | 0.000 | 0.000 | 0.202 | 0.131 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-3.5.150 | 0.000 | 0.000 | 0.215 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4.30 | 8.968 | 6.771 | 10.381 | 9.229 | 7.785 | 8.362 | 8.953 | 9.167 | 8.319 |
| GPT-4.45 | 7.812 | 8.734 | 9.059 | 6.778 | 8.171 | 7.692 | 7.106 | 7.350 | 6.969 |
| GPT-4.60 | 6.952 | 7.509 | 7.664 | 7.748 | 7.664 | 8.456 | 8.007 | 8.070 | 7.692 |
| GPT-4.90 | 7.533 | 10.078 | 8.687 | 8.920 | 8.494 | 6.809 | 8.083 | 6.852 | 7.940 |
| GPT-4.120 | 8.982 | 10.288 | 10.656 | 10.183 | 7.787 | 5.992 | 8.367 | 8.039 | 7.615 |
| GPT-4.150 | 8.917 | 10.088 | 10.917 | 9.239 | 7.642 | 6.388 | 8.051 | 7.083 | 7.249 |
| GPT-4o.30 | 6.768 | 6.424 | 8.478 | 6.893 | 5.709 | 5.401 | 7.095 | 6.500 | 6.282 |
| GPT-4o.45 | 8.333 | 8.021 | 10.835 | 7.015 | 7.282 | 7.692 | 9.012 | 7.521 | 8.014 |
| GPT-4o.60 | 8.734 | 8.608 | 10.766 | 7.869 | 7.299 | 8.272 | 9.431 | 7.719 | 8.766 |
| GPT-4o.90 | 8.286 | 10.271 | 9.459 | 8.291 | 8.687 | 7.393 | 9.586 | 7.963 | 8.696 |
| GPT-4o.120 | 7.784 | 8.848 | 10.861 | 9.008 | 6.557 | 7.231 | 8.964 | 7.843 | 7.415 |
| GPT-4o.150 | 8.280 | 8.333 | 10.480 | 9.783 | 7.205 | 5.507 | 8.263 | 7.292 | 7.463 |
| GARCH.LPA | 1.289 | 1.705 | 2.075 | 1.453 | 0.566 | 0.947 | 1.287 | 1.268 | 1.664 |
| EWMA.N.80 | 1.636 | 2.056 | 2.048 | 2.570 | 1.304 | 1.121 | 1.815 | 1.610 | 1.825 |
| EWMA.N.120 | 0.980 | 1.818 | 1.811 | 2.570 | 1.610 | 1.616 | 1.370 | 1.541 | 1.575 |
| EWMA.DCS.N.80 | 0.909 | 0.374 | 0.559 | 0.670 | 0.559 | 0.187 | 0.907 | 0.894 | 0.547 |
| EWMA.DCS.N.120 | 0.784 | 0.606 | 0.402 | 0.670 | 0.604 | 0.202 | 0.783 | 0.963 | 0.591 |
| EWMA.DCS.T.80 | 0.182 | 0.187 | 0.559 | 0.670 | 0.000 | 0.000 | 0.181 | 0.000 | 0.000 |
| EWMA.DCS.T.120 | 0.000 | 0.000 | 0.402 | 0.670 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GARCH.N.120 | 0.784 | 1.414 | 1.610 | 2.346 | 1.408 | 1.818 | 0.978 | 1.734 | 0.787 |
| GARCH.N.250 | 1.053 | 0.000 | 1.090 | 1.899 | 0.272 | 0.000 | 1.050 | 0.771 | 1.058 |
| GARCH.DCS.N.120 | 0.784 | 1.616 | 1.006 | 2.346 | 1.207 | 1.212 | 0.978 | 1.734 | 0.787 |
| GARCH.DCS.N.250 | 1.053 | 0.000 | 1.090 | 1.788 | 0.272 | 0.000 | 1.050 | 0.771 | 1.058 |
| GARCH.DCS.T.120 | 0.588 | 0.000 | 0.604 | 1.229 | 0.402 | 0.000 | 0.196 | 0.385 | 0.394 |
| GARCH.DCS.T.250 | 0.526 | 0.000 | 0.272 | 1.229 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

*Note:* 1. Green: POF below 1%, 2. Red: POF above 1%.

Table 2 presents the failure rates, showing occurrences where losses exceeded the 1% VaR threshold, and Table B1 from Appendix B provides

Kupiec's POF test p-values, with green indicating well-calibrated models and red highlighting underperforming ones.

Among LLMs, GPT-3.5 shows the best calibration, particularly in the 30-day window, with failure rates close to 1% for indices like CRIX and SPGT-CLTR. As the window extends from 45 to 150 days, GPT-3.5 approaches 0% exceedance rate for all indices, suggesting potentially conservative risk estimation.

In contrast, GPT-4 and GPT-4o consistently fail to capture risk accurately, displaying failure rates well above 1%, with the 30-day window for GPT-4 reaching as high as 10.38% for FTSE.

Classical models, such as GARCH-LPA and EWMA-DCS, demonstrate greater reliability, particularly in the Kupiec test. GARCH-LPA has moderate failure rates, such as 1.29% for CACT, while EWMA-DCS—especially the normal innovation variant—provides the most consistent performance, with EWMA-DCS.N.120 achieving failure rates near the 1% target (0.91% for CACT and 0.67% for CRIX) and Kupiect test p-values close to 1, underscoring robust risk estimation.

The Kupiec test confirms GPT-3.5's superior calibration among LLMs in the 30-day window, with p-values near 1 across most indices. However, GPT-3.5's calibration deteriorates as the rolling window increases, while GPT-4 and GPT-4o continue to show significant deviations across all indices.

Table 3: Traffic light test for 1% VaR: $\Phi(z)$.

| Model | CACT | DJCI | FTSE | CRIX | CBU | SP500 | STOXX | SPGTCLTR | GDAXI |
|---|---|---|---|---|---|---|---|---|---|
| GPT-3.5.30 | 0.515 | 0.540 | 0.904 | 0.424 | 0.523 | 0.378 | 0.513 | 0.109 | 0.518 |
| GPT-3.5.45 | 0.124 | 0.063 | 0.127 | 0.202 | 0.024 | 0.009 | 0.123 | 0.055 | 0.125 |
| GPT-3.5.60 | 0.025 | 0.028 | 0.138 | 0.033 | 0.026 | 0.010 | 0.009 | 0.060 | 0.009 |
| GPT-3.5.90 | 0.010 | 0.011 | 0.077 | 0.007 | 0.011 | 0.011 | 0.010 | 0.010 | 0.010 |
| GPT-3.5.120 | 0.012 | 0.013 | 0.037 | 0.008 | 0.013 | 0.014 | 0.012 | 0.012 | 0.012 |
| GPT-3.5.150 | 0.015 | 0.016 | 0.044 | 0.003 | 0.016 | 0.016 | 0.014 | 0.014 | 0.015 |
| GPT-4.30 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4.45 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4.60 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4.90 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4.120 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4.150 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4o.30 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4o.45 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4o.60 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4o.90 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4o.120 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4o.150 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GARCH.LPA | 0.751 | 0.948 | 0.994 | 0.913 | 0.158 | 0.451 | 0.749 | 0.737 | 0.940 |
| EWMA.N.80 | 0.933 | 0.993 | 0.993 | 1.000 | 0.760 | 0.611 | 0.973 | 0.926 | 0.974 |
| EWMA.N.120 | 0.482 | 0.966 | 0.965 | 1.000 | 0.914 | 0.916 | 0.800 | 0.892 | 0.904 |
| EWMA.DCS.N.80 | 0.415 | 0.073 | 0.152 | 0.161 | 0.152 | 0.029 | 0.414 | 0.401 | 0.143 |
| EWMA.DCS.N.120 | 0.312 | 0.189 | 0.090 | 0.161 | 0.187 | 0.037 | 0.311 | 0.467 | 0.177 |
| EWMA.DCS.T.80 | 0.027 | 0.029 | 0.152 | 0.161 | 0.010 | 0.010 | 0.027 | 0.009 | 0.009 |
| EWMA.DCS.T.120 | 0.012 | 0.013 | 0.090 | 0.161 | 0.013 | 0.013 | 0.012 | 0.011 | 0.012 |
| GARCH.N.120 | 0.312 | 0.823 | 0.914 | 1.000 | 0.820 | 0.966 | 0.480 | 0.954 | 0.315 |
| GARCH.N.250 | 0.541 | 0.027 | 0.569 | 0.997 | 0.081 | 0.027 | 0.539 | 0.325 | 0.545 |
| GARCH.DCS.N.120 | 0.312 | 0.916 | 0.505 | 1.000 | 0.679 | 0.682 | 0.480 | 0.954 | 0.315 |
| GARCH.DCS.N.250 | 0.541 | 0.027 | 0.569 | 0.991 | 0.081 | 0.028 | 0.539 | 0.325 | 0.545 |
| GARCH.DCS.T.120 | 0.175 | 0.013 | 0.187 | 0.754 | 0.090 | 0.013 | 0.034 | 0.080 | 0.085 |
| GARCH.DCS.T.250 | 0.177 | 0.027 | 0.081 | 0.754 | 0.027 | 0.027 | 0.025 | 0.024 | 0.025 |

*Note:* 1. Green: accurate risk estimation, 2. Yellow: some uncertainty, 3. Red: potential inaccuracies.

Table 3 presents the Traffic Light Test results for the 1% VaR across different indices. The color scheme highlights model performance: green indicates accurate risk estimation, yellow suggests some uncertainty, and red denotes potential inaccuracies. The GPT-3.5 model demonstrates outstanding accuracy across all windows and assets, suggesting accurate tail risk estimation. In contrast, the GPT-4 and GPT-4o models perform poorly across all windows, with $\Phi(z)$ values at 1 across all indices, indicating severe risk misestimation. Among the classical models, GARCH and EWMA-DCS models excel, with green cells for all assets.

Table 4: Christoffersen Conditional Coverage test: p-values.

| Model | CACT | DJCI | FTSE | CRIX | CBU | SP500 | STOXX | SPGTCLTR | GDAXI |
|---|---|---|---|---|---|---|---|---|---|
| GPT-3.5.30 | 0.939 | 0.934 | 0.010 | 0.166 | 0.938 | 0.825 | 0.940 | 0.391 | 0.939 |
| GPT-3.5.45 | 0.440 | 0.210 | 0.451 | 0.065 | 0.051 | 0.001 | 0.438 | 0.180 | 0.444 |
| GPT-3.5.60 | 0.055 | 0.062 | 0.484 | 0.106 | 0.057 | 0.001 | 0.001 | 0.198 | 0.001 |
| GPT-3.5.90 | 0.001 | 0.001 | 0.262 | 0.007 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| GPT-3.5.120 | 0.002 | 0.002 | 0.094 | 0.010 | 0.002 | 0.002 | 0.001 | 0.001 | 0.002 |
| GPT-3.5.150 | 0.002 | 0.002 | 0.119 | 0.000 | 0.002 | 0.003 | 0.002 | 0.002 | 0.002 |
| GPT-4.30 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4.45 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4.60 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4.90 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4.120 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4.150 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.30 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.45 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.60 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.90 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.120 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.150 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GARCH.LPA | 0.737 | 0.285 | 0.043 | 0.178 | 0.543 | 0.946 | 0.740 | 0.758 | 0.313 |
| EWMA.N.80 | 0.333 | 0.078 | 0.046 | 0.000 | 0.723 | 0.898 | 0.186 | 0.339 | 0.181 |
| EWMA.N.120 | 0.951 | 0.218 | 0.091 | 0.000 | 0.397 | 0.391 | 0.659 | 0.454 | 0.425 |
| EWMA.DCS.N.80 | 0.934 | 0.248 | 0.527 | 0.553 | 0.527 | 0.068 | 0.933 | 0.839 | 0.501 |
| EWMA.DCS.N.120 | 0.853 | 0.628 | 0.313 | 0.553 | 0.623 | 0.094 | 0.851 | 0.949 | 0.596 |
| EWMA.DCS.T.80 | 0.060 | 0.068 | 0.527 | 0.051 | 0.001 | 0.001 | 0.060 | 0.001 | 0.001 |
| EWMA.DCS.T.120 | 0.001 | 0.002 | 0.313 | 0.051 | 0.002 | 0.002 | 0.001 | 0.001 | 0.001 |
| GARCH.N.120 | 0.853 | 0.616 | 0.126 | 0.002 | 0.621 | 0.218 | 0.951 | 0.104 | 0.857 |
| GARCH.N.250 | 0.953 | 0.007 | 0.942 | 0.034 | 0.253 | 0.007 | 0.953 | 0.876 | 0.951 |
| GARCH.DCS.N.120 | 0.853 | 0.391 | 0.950 | 0.002 | 0.838 | 0.834 | 0.951 | 0.266 | 0.857 |
| GARCH.DCS.N.250 | 0.953 | 0.007 | 0.942 | 0.058 | 0.253 | 0.007 | 0.953 | 0.876 | 0.951 |
| GARCH.DCS.T.120 | 0.591 | 0.002 | 0.623 | 0.239 | 0.313 | 0.002 | 0.083 | 0.274 | 0.293 |
| GARCH.DCS.T.250 | 0.591 | 0.007 | 0.253 | 0.239 | 0.007 | 0.007 | 0.006 | 0.005 | 0.006 |

*Note:* 1. Green: p-values above 0.05. 2. Red: p-values below 0.05.

Table 4 presents p-values from Christoffersen's Conditional Coverage test, assessing the independence of exceedances. GPT-3.5 models with 30- and 45-day windows perform well, with high p-values across most assets, indicating effective capture of both the frequency and independence of risk events. However, for longer backtesting windows (60 up to 150 days), GPT-3.5 tends to overestimate risk, leading to low p-values that suggest a lack of independence in exceedances.

GPT-4 and GPT-4o models fail to capture risk effectively, displaying consistently low p-values across all windows and assets, indicating issues with both the frequency and independence of risk events, as highlighted in Appendix B, Figures B5 and B6.

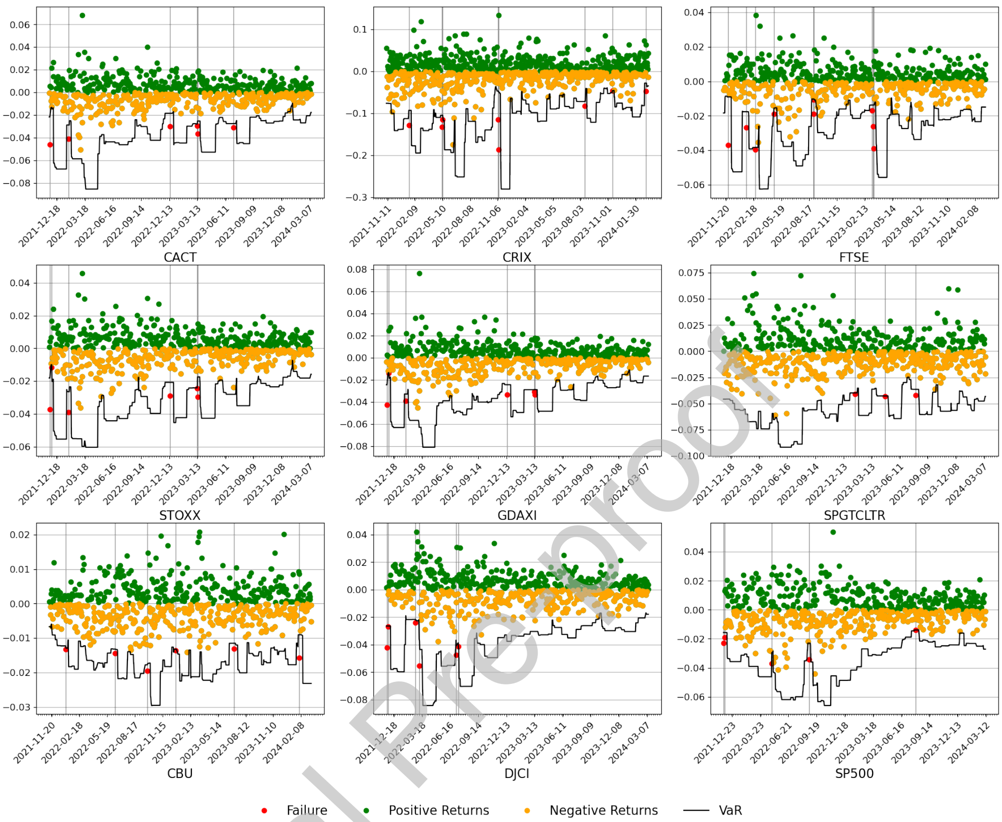



Figure 3: VaR Exceedances for LLM-VaR GPT-3.5 (30-day rolling window).

*Note:* Red dots indicate exceedances. Green dots show gains, and Orange dots show losses within the expected range.



GARCH models perform well in stable markets but struggle in more volatile settings, particularly GARCH-N.120 and GARCH-N.250. In contrast, EWMA models, especially EWMA-DCS with normal innovations, exhibit robust performance across various market conditions, reliably capturing risk independence.

GPT-3.5 with a 30-day and 45-day window, along with EWMA models and GARCH-LPA, emerge as strong performers in predicting 1% VaR, as illustrated in Figure 3 and Figures B3 and B4, from Appendix B.

### *4.2.2. ES Backtesting Results*

To address the limitations of VaR in capturing tail risk beyond a specific quantile, we apply $Z_2$ and $Z_3$ tests for ES (Acerbi and Székely, 2014). The former tests for the frequency and magnitude of ES violations, while the latter tests for the independence of exceedances. We choose to present test statistics here (partially because of the well-known $Z_2^*$ threshold), but colors are still assigned based on simulated p-values.

Table 5: $Z_2$ test statistic.

| Model | CACT | DJCI | FTSE | CRIX | CBU | SP500 | STOXX | SPGTCLTR | GDAXI |
|---|---|---|---|---|---|---|---|---|---|
| GPT-3.5.30 | 0.406 | 0.471 | -0.140 | 0.348 | 0.510 | 0.552 | 0.406 | 0.788 | 0.489 |
| GPT-3.5.45 | 0.763 | 0.812 | 0.659 | 0.621 | 0.924 | 1.000 | 0.763 | 0.853 | 0.772 |
| GPT-3.5.60 | 0.920 | 0.889 | 0.691 | 0.828 | 0.922 | 1.000 | 1.000 | 0.853 | 1.000 |
| GPT-3.5.90 | 1.000 | 1.000 | 0.773 | 0.943 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-3.5.120 | 1.000 | 1.000 | 0.897 | 0.943 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-3.5.150 | 1.000 | 1.000 | 0.899 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GPT-4.30 | -9.838 | -3.555 | -7.907 | -5.997 | -4.084 | -4.625 | 16.495 | -4.190 | 1.564 |
| GPT-4.45 | -3.832 | -4.002 | -4.791 | -3.893 | -3.311 | -3.135 | -3.418 | -3.328 | -3.540 |
| GPT-4.60 | -3.353 | -3.141 | -4.633 | -4.602 | -3.392 | -3.337 | -3.916 | -4.234 | -4.191 |
| GPT-4.90 | -3.626 | -5.036 | -4.913 | -5.085 | -3.502 | -2.618 | -4.131 | -3.175 | -3.991 |
| GPT-4.120 | -4.596 | -5.021 | -5.863 | -6.645 | 10.288 | -2.623 | -4.191 | -3.534 | -4.143 |
| GPT-4.150 | -4.351 | 5.312 | -6.932 | -6.744 | -3.352 | -2.498 | -5.139 | -3.055 | -3.482 |
| GPT-4o.30 | 17.059 | -2.881 | -5.050 | -4.361 | -2.188 | -2.261 | -8.009 | -2.464 | -4.348 |
| GPT-4o.45 | -4.023 | -3.485 | -6.117 | -4.599 | -2.964 | -3.069 | -4.539 | -3.204 | -4.135 |
| GPT-4o.60 | -4.700 | -4.325 | -6.102 | -4.832 | -2.978 | -3.389 | -4.962 | -3.462 | -4.604 |
| GPT-4o.90 | -4.251 | -5.257 | -5.687 | -5.398 | -3.629 | -2.823 | -4.655 | -3.451 | -4.452 |
| GPT-4o.120 | -3.782 | -3.989 | -6.152 | -5.898 | -2.556 | -2.812 | -4.106 | -3.431 | -3.354 |
| GPT-4o.150 | -3.959 | -3.800 | -5.796 | -5.957 | -2.776 | -1.925 | -3.759 | -3.132 | -3.526 |
| GARCH.LPA | 0.340 | 0.228 | 0.023 | 0.222 | 0.728 | 0.554 | 0.384 | 0.296 | 0.162 |
| EWMA.N.80 | 0.054 | 0.037 | -0.169 | -0.440 | 0.369 | 0.485 | 0.013 | 0.152 | 0.076 |
| EWMA.N.120 | 0.405 | 0.151 | -0.019 | -0.440 | 0.227 | 0.267 | 0.247 | 0.143 | 0.188 |
| EWMA.DCS.N.80 | 0.527 | 0.844 | 0.684 | 0.616 | 0.763 | 0.923 | 0.566 | 0.577 | 0.738 |
| EWMA.DCS.N.120 | 0.609 | 0.748 | 0.799 | 0.616 | 0.744 | 0.917 | 0.637 | 0.514 | 0.718 |
| EWMA.DCS.T.80 | 0.923 | 0.927 | 0.723 | 0.660 | 1.000 | 1.000 | 0.920 | 1.000 | 1.000 |
| EWMA.DCS.T.120 | 1.000 | 1.000 | 0.800 | 0.660 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GARCH.N.120 | 0.569 | 0.353 | 0.165 | -0.282 | 0.346 | 0.179 | 0.513 | 0.165 | 0.589 |
| GARCH.N.250 | 0.504 | 1.000 | 0.468 | -0.019 | 0.874 | 1.000 | 0.530 | 0.634 | 0.539 |
| GARCH.DCS.N.120 | 0.585 | 0.282 | 0.508 | -0.278 | 0.451 | 0.461 | 0.515 | 0.160 | 0.591 |
| GARCH.DCS.N.250 | 0.498 | 1.000 | 0.473 | 0.030 | 0.878 | 1.000 | 0.530 | 0.634 | 0.538 |
| GARCH.DCS.T.120 | 0.715 | 1.000 | 0.688 | 0.365 | 0.833 | 1.000 | 0.922 | 0.798 | 0.811 |
| GARCH.DCS.T.250 | 0.789 | 1.000 | 0.872 | 0.394 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

*Note:* 1. Green: p-values above 0.05. 2. Red: p-values below 0.05.

Table 5 presents the $Z_2$ test results, indicating that GPT-4 and GPT-4o yield markedly negative statistics, falling well below the critical threshold $Z_2^*$. This is expected, given the poor VaR calibration presented earlier, which extends to the predicted distribution as a whole. The small percentage of index-window size configurations where we see large statistics are outliers due to a few miss-specified ES samples (very close to 0), which produce large positive statistic terms and outweigh all other VaR breaches with negative contributions. On the other hand, GPT-3.5 can obtain solid scores for all windows. We suppose that its tendency to predict the left side of the return distribution secures reasonable VaR and ES calibration. Regarding benchmark models, it is clear that most pass the test. EWMA with normal residuals struggles more than other methods, and CRIX appears to be more challenging to fit for GARCH with normal residuals (also when including DCS).

Table 6: $Z_3$ test statistic.

| Model | CACT | DJCI | FTSE | CRIX | CBU | SP500 | STOXX | SPGTCLTR | GDAXI |
|---|---|---|---|---|---|---|---|---|---|
| GPT-3.5.30 | 0.068 | 0.094 | 0.020 | 0.064 | 0.080 | 0.100 | 0.066 | 0.108 | 0.050 |
| GPT-3.5.45 | 0.104 | 0.123 | 0.096 | 0.088 | 0.137 | 0.174 | 0.112 | 0.127 | 0.116 |
| GPT-3.5.60 | 0.144 | 0.144 | 0.106 | 0.118 | 0.125 | 0.170 | 0.148 | 0.137 | 0.140 |
| GPT-3.5.90 | 0.155 | 0.142 | 0.115 | 0.125 | 0.163 | 0.166 | 0.157 | 0.152 | 0.157 |
| GPT-3.5.120 | 0.162 | 0.160 | 0.144 | 0.148 | 0.152 | 0.170 | 0.161 | 0.171 | 0.150 |
| GPT-3.5.150 | 0.156 | 0.161 | 0.139 | 0.174 | 0.150 | 0.172 | 0.166 | 0.150 | 0.168 |
| GPT-4.30 | -0.572 | -0.312 | -0.477 | -0.613 | -0.326 | -0.384 | -0.476 | -0.307 | -0.421 |
| GPT-4.45 | -0.528 | -0.526 | -0.567 | -0.738 | -1.444 | -0.343 | -0.498 | -0.452 | -0.528 |
| GPT-4.60 | -0.572 | -0.473 | -0.648 | -0.836 | -0.442 | -0.363 | -0.551 | -0.592 | -0.573 |
| GPT-4.90 | -0.589 | 1.432 | -0.623 | -0.928 | -0.425 | -0.409 | -0.564 | -0.500 | -0.601 |
| GPT-4.120 | -0.596 | -0.498 | -0.745 | -0.998 | -0.353 | -0.606 | -0.578 | -0.455 | -0.631 |
| GPT-4.150 | -0.591 | -0.345 | -0.672 | -0.955 | -0.441 | -0.443 | -0.588 | -0.484 | -0.661 |
| GPT-4o.30 | -0.439 | -0.463 | -0.502 | -0.534 | -0.352 | -0.368 | -0.389 | -0.404 | -0.483 |
| GPT-4o.45 | -0.537 | -0.505 | -0.639 | -0.691 | -0.423 | -0.381 | -0.524 | -0.461 | -0.549 |
| GPT-4o.60 | -0.590 | -0.590 | -0.663 | -0.835 | -0.445 | -0.462 | -0.566 | -0.539 | -0.644 |
| GPT-4o.90 | -0.591 | -0.568 | -0.682 | -0.866 | -0.438 | -0.447 | -0.607 | -0.531 | -0.616 |
| GPT-4o.120 | -0.626 | -0.551 | -0.742 | -0.870 | -0.433 | -0.480 | -0.598 | -0.499 | -0.521 |
| GPT-4o.150 | -0.607 | -0.522 | -0.691 | -0.862 | -0.421 | -0.418 | -0.554 | -0.497 | -0.564 |
| GARCH.LPA | 0.989 | 0.989 | 0.991 | 0.959 | 0.995 | 0.988 | 0.991 | 0.982 | 0.988 |
| EWMA.N.80 | -0.364 | -0.214 | -0.445 | -0.509 | -0.141 | -0.120 | -0.341 | -0.262 | -0.266 |
| EWMA.N.120 | -0.247 | -0.190 | -0.384 | -0.509 | -0.173 | -0.146 | -0.255 | -0.303 | -0.236 |
| EWMA.DCS.N.80 | -0.032 | 0.062 | -0.048 | -0.056 | 0.101 | 0.120 | -0.032 | 0.005 | 0.063 |
| EWMA.DCS.N.120 | 0.023 | 0.067 | 0.011 | -0.056 | 0.101 | 0.110 | 0.015 | -0.024 | 0.049 |
| EWMA.DCS.T.80 | 0.174 | 0.270 | 0.092 | 0.036 | 0.316 | 0.340 | 0.201 | 0.252 | 0.228 |
| EWMA.DCS.T.120 | 0.285 | 0.289 | 0.148 | 0.036 | 0.296 | 0.314 | 0.278 | 0.235 | 0.278 |
| GARCH.N.120 | -0.068 | -0.131 | -0.234 | -0.434 | -0.119 | -0.151 | -0.071 | -0.190 | -0.076 |
| GARCH.N.250 | -0.021 | 0.128 | -0.113 | -0.333 | 0.034 | 0.200 | 0.022 | -0.073 | 0.035 |
| GARCH.DCS.N.120 | -0.043 | -0.119 | -0.100 | -0.431 | -0.088 | -0.098 | -0.061 | -0.187 | -0.082 |
| GARCH.DCS.N.250 | -0.026 | 0.127 | -0.108 | -0.325 | 0.073 | 0.201 | 0.024 | -0.076 | 0.037 |
| GARCH.DCS.T.120 | 0.115 | 0.228 | 0.021 | -0.168 | 0.129 | 0.236 | 0.199 | 0.044 | 0.163 |
| GARCH.DCS.T.250 | 0.269 | 0.394 | 0.166 | -0.109 | 0.288 | 0.434 | 0.283 | 0.212 | 0.320 |

*Note:* 1. Green: p-values above 0.05. 2. Red: p-values below 0.05.

The results of $Z_3$ test are reported in Table 6. This time, we test for exceedance ranks and estimate ES empirically. Therefore, this test does not involve VaR or ES estimations. A similar story emerges: we find that EWMA.N does not provide independent exceedances, with more failing indices than $Z_2$. In a similar vein, the GARCH.N models do not produce independent exceedances for CRIX.

For a clearer picture of how LLMs' predictions compare to actual log returns, Appendix C presents distribution plots (Figures C1, C2, C3).

### *4.3. Sensitivity Analysis*

All inference parameters have good defaults according to the experiments in Gruver et al. (2024), which other works have silently adopted (Cao and

Wang, 2024; Tang et al., 2025).

Due to its impact on model creativity, the temperature parameter $\tau$ prompted a separate ablation study. We expect it to be the most sensitive to changes.

Table 7: Temperature sensitivity analysis parameter space.

| Parameter | Values |
|---|---|
| LLM | {GPT-3.5, GPT-4, GPT-4o} |
| Asset | CRIX |
| $\tau$ | $\{0.0, 0.1, 0.2, \ldots, 0.9, 1.0\}$ |
| $\omega$ | 45 |
| $\alpha_{LLM}$ | 0.95 |
| $\beta_{LLM}$ | 0.35 |
| $\pi$ | 2 |

To test the influence of the temperature parameter, we performed a series of experiments using the setup described in Table 7. The other parameters were chosen to align as closely as possible with those used in the referenced paper. CRIX, as a representative risky index, provides a reasonable testbed for this analysis.

Table 8: Temperature sensitivity analysis backtesting results. VaR, ES presents the number of passing VaR and ES backtests, POF the failure rate.

| | GPT-3.5 | | | GPT-4 | | | GPT-4o | | |
|---|---|---|---|---|---|---|---|---|---|
| **Temperature** | **VaR** | **ES** | **POF** | **VaR** | **ES** | **POF** | **VaR** | **ES** | **POF** |
| 0.0 | 3 | 1 | 0.71% | 0 | 1 | 16.41% | 0 | 0 | 16.88% |
| 0.1 | 3 | 1 | 0.71% | 0 | 0 | 13.08% | 0 | 0 | 9.75% |
| 0.2 | 3 | 1 | 0.71% | 0 | 0 | 11.30% | 0 | 0 | 8.68% |
| 0.3 | 3 | 2 | 0.71% | 0 | 0 | 9.16% | 0 | 0 | 8.56% |
| 0.4 | 3 | 2 | 0.71% | 0 | 0 | 8.68% | 0 | 0 | 7.73% |
| 0.5 | 3 | 2 | 0.71% | 0 | 0 | 7.85% | 0 | 0 | 7.37% |
| 0.6 | 3 | 2 | 0.71% | 0 | 0 | 7.73% | 0 | 0 | 7.49% |
| 0.7 | 3 | 2 | 0.71% | 0 | 0 | 6.30% | 0 | 0 | 7.25% |
| 0.8 | 3 | 2 | 0.71% | 0 | 0 | 5.95% | 0 | 0 | 6.90% |
| 0.9 | 3 | 2 | 0.71% | 0 | 0 | 5.35% | 0 | 0 | 6.54% |
| 1.0 | 3 | 2 | 0.71% | 0 | 0 | 4.64% | 0 | 0 | 6.06% |

*Note:* 1. Green: Maximum number of tests pass. 2. Red: No test passes.

We summarize our main findings in Table 8. Passing VaR and ES backtests are counted separately for each combination of parameters. We define "passing" for VaR as: p-value $> 0.05$ for the Kupiec POF and Christoffersen tests, value $< 0.95$ for the traffic light test. Consequently, the ES $Z_2$ and $Z_3$ tests must yield p-values $> 0.05$ to be considered successful. The failure rate is also shown for comparison purposes.

We note that GPT-3.5 is stable across different temperatures, with similar failure rates and VaR prediction performance. The $Z_3$ test for ES is "passed" only starting from higher temperatures, with increasing confidence (not shown here), which could indicate dependencies for values under the tail, a phenomenon exacerbated by the almost uniform predictions offered for small $\tau$. On the other hand, GPT-4 and 4o steadily improve their failure rates for higher temperatures, which is an interesting finding. We suspect that affording more creative liberty tends to output returns closer to the tails, although this is an avenue for further studies. However, VaR and ES are not calibrated well enough to produce satisfactory backtesting performance. Even more, it is not recommended to go above $\tau = 1.0$ (OpenAI, 2025b).

We conclude that the temperature parameter does not significantly influence our results. Therefore, we retain $\tau = 0.7$, as recommended by Gruver et al. (2024).

## 5. Limitations

### 5.1. Inference Costs

In our setup, a notable limitation of using LLM-based models for risk estimation is the cost associated with each forecasting day. Since we rely on paid models, their expenses can accumulate, particularly in applications requiring frequent or high-volume predictions. Table 9 presents the LLM costs per forecasting day per asset [6] for different LLMs in our configuration, illustrating that more advanced models like GPT-4 incur higher costs. We consider the benchmarks cost-free.

[6] Costs valid for June 2025.

Table 9: LLM costs per forecasting day/asset

| **Model** | **Cost (USD)** |
|---|---|
| GPT-3.5 | 0.010 |
| GPT-4 | 0.071 |
| GPT-4o | 0.037 |

Runtime is an important factor that contributes to timely management decisions. Mean daily runtimes and their standard deviation are presented in Table 10. GPT-3.5 performs closer to the benchmarks, which deliver instantaneous results, except for GARCH-LPA. Larger LLMs exhibit increased runtimes, with as much as 10 - 15 seconds of variability. Waiting times thus increase at a higher rate for historical forecasts.

Table 10: Mean and standard deviations of daily runtimes for methods.

| **Model** | **Mean** | **Standard deviation** |
|---|---|---|
| GPT-3.5.30 | 1.56 | 1.66 |
| GPT-3.5.45 | 1.46 | 0.55 |
| GPT-3.5.60 | 2.42 | 0.46 |
| GPT-3.5.90 | 2.47 | 0.67 |
| GPT-3.5.120 | 2.56 | 0.84 |
| GPT-3.5.150 | 2.57 | 1.17 |
| GPT-4.30 | 21.77 | 10.79 |
| GPT-4.45 | 28.04 | 11.35 |
| GPT-4.60 | 26.46 | 9.87 |
| GPT-4.90 | 28.52 | 14.19 |
| GPT-4.120 | 27.50 | 10.93 |
| GPT-4.150 | 24.70 | 13.41 |
| GPT-4o.30 | 21.17 | 7.94 |
| GPT-4o.45 | 24.69 | 8.51 |
| GPT-4o.60 | 27.73 | 9.48 |
| GPT-4o.90 | 25.75 | 12.40 |
| GPT-4o.120 | 28.30 | 10.57 |
| GPT-4o.150 | 26.50 | 11.46 |
| GARCH.LPA | 2.72 | 0.52 |
| EWMA.DCS.N.120 | 0.00 | 0.00 |
| EWMA.DCS.N.80 | 0.00 | 0.00 |
| EWMA.N.120 | 0.00 | 0.00 |
| EWMA.N.80 | 0.00 | 0.00 |
| EWMA.DCS.T.120 | 0.00 | 0.00 |
| EWMA.DCS.T.80 | 0.00 | 0.00 |
| GARCH.GAS.N.120 | 0.20 | 0.07 |
| GARCH.GAS.N.250 | 0.39 | 0.11 |
| GARCH.GAS.T.120 | 0.22 | 0.07 |
| GARCH.GAS.T.250 | 0.41 | 0.12 |
| GARCH.N.120 | 0.13 | 0.05 |
| GARCH.N.250 | 0.28 | 0.09 |

*Note:* LLMs are benchmarked on the last 30 days for CRIX. Traditional methods are computed on the full out-of-sample dataset and averaged for all assets. All times are in seconds.

This introduces a trade-off between cost and flexibility: LLMs require no parameter tuning and support zero-shot adaptability across assets and tasks, while traditional models offer cost-efficiency, fast runtimes, transparency, and established regulatory acceptance. The viability of LLM-based forecasting thus depends on institutional priorities such as scalability, explainability, and responsiveness to market changes versus infrastructure and budget constraints, similar to the arguments of Li et al. (2023).

### *5.2. Performance, Data Privacy and Model Availability*

LLM-VaR and LLM-ES, estimated using GPT-3.5 through the LLMTime framework, exhibit strong performance for shorter rolling windows (30 and 45 days). However, for longer windows, GPT-3.5 tends to generate conservative estimates, resulting in overly cautious risk forecasts (see Figures B1 and B2, from Appendix B). In such cases, traditional models like GARCH—designed to capture persistent volatility patterns—often provide more reliable long-term forecasts.

Beyond the length of the rolling window, the effectiveness of LLM-VaR and LLM-ES also depends on the quality of historical data. In markets with sparse or noisy signals, LLM performance may degrade. While general-purpose models offer adaptability, their broad training objectives may limit precision in domain-specific tasks without fine-tuning.

Irrespective of model performance, a principal limitation of LLMs concerns their black-box nature, impeding interpretability and posing significant challenges in regulatory environments that require transparency and model validation, particularly during periods of market stress (Li et al., 2023). While a growing body of work on explainable AI (XAI) for time series models (e.g., Bento et al., 2021) provides tools to shed some light on the LLM-induced mapping of input data to forecasts, the characteristic approach of passing model inputs via a prompt and associated degrees of freedom can easily jeopardize standard XAI time series approaches and demands LLM-specific solutions.

Relatedly, interfacing LLMs via commercial APIs raises concerns about data privacy and sustainability, as model providers could decide to discontinue access to a model (version). In financial contexts, regulatory constraints often prohibit external processing of sensitive or proprietary time series. Although our study only uses publicly available data, real-world applications would require secure, on-premises deployment or privacy-preserving inference mechanisms. One one hand, OpenAI's data privacy policy ensures that no information fed to their paid APIs will be used for model training (OpenAI, 2025a), and other vendors offer similar contracts. On the other hand, the advent of powerful open-weight LLMs, such as Meta's LLaMA 3.1(Touvron et al., 2024), Google's Gemma 2 (Google DeepMind, 2024), Mistral Large 2 (Mistral AI, 2024), or more recently Qwen3 (Yang et al., 2025) and DeepSeek (DeepSeek-AI et al., 2025), facilitates mitigating privacy and sustainability risks through the deployment of on-premise LLM-based forecasting solutions.

A specific limitation of this study's setup may be seen in the exclusive reliance on OpenAI's GPT models. Being the first study of its kind, we favored this setup because it facilitated controlled comparisons across model generations (3.5, 4, and 4o). However, we acknowledge that our focus on GPT-type LLMs restricts generalizability, calling for future work to evaluate alternative LLM ecosystems—such as Anthropic's Claude, Google's Gemini, to name a few, which may exhibit different alignment behaviors, numerical stability, and domain generalization capacities. Specifically, we observe GPT-4 and GPT-4o to perform inferior to GPT-3.5 in risk estimation tasks, suggesting that improvements in general language modeling do not necessarily translate into better quantitative forecasting. We deem this phenomenon worthy of further investigation and attempt to provide some answers in the next section. In general, studying different LLM variants can potentially uncover the architectural patterns that govern a model's adequacy for forecasting and/or risk management.

## 6. Discussion

### *6.1. Poor Performance of Newer Models*

A notable finding from our results is that GPT-3.5 outperforms GPT-4 and GPT-4o in forecasting VaR and ES. This is somewhat counterintuitive, given that the latter models are newer and trained on broader, more diverse datasets. Several plausible explanations may account for this outcome.

First, the reduced performance in the case of GPT-4 and GPT-4o can be to some extent attributed to Reinforcement Learning from Human Feedback (RLHF), a key element in their fine-tuning, which introduces a further layer of alignment, as noticed by Gruver et al. (2024). Although RLHF improves safety and helps generate responses more in line with human preferences, it additionally biases the model toward overconfidence. This can be seen by comparing the alignment of responses and the expected answer probabilities for the MMLU dataset in the case of GPT-4 (OpenAI, 2023). In contrast, GPT-3.5, which lacks RLHF fine-tuning, does not exhibit the same behavior.

Second, we can explain the good tail predictions of GPT-3.5 by its strong bias for the left tail, as opposed to its newer variants. Essentially, while forecasting the whole return distribution with LLMTime, GPT-3.5 is more concerned with the left tail. We can see this visually in Appendix C. Further confirmation is achieved by evaluating the return forecast performance, not VaR or ES. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are computed and averaged for all assets in Table 11. One notices that GPT-4 and 4o exhibit better performance, a further indication that they focus on the entire distribution, not only the left tail. Additionally, when tested on standard benchmark datasets, some with more obvious seasonality and trend patterns, ARIMA has been reported to outperform GPT-3.5 (Cao and Wang, 2024), although opinions are mixed (Tang et al., 2025).

Table 11: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) averaged over all assets, for each LLM and window size

| | **GPT-3.5** | | **GPT-4** | | **GPT-4o** | |
|---|---|---|---|---|---|---|
| **Window size** | **MAE** | **RMSE** | **MAE** | **RMSE** | **MAE** | **RMSE** |
| 30 | 0.0435 | 0.0489 | 0.0106 | 0.0146 | ***0.0099*** | ***0.0139*** |
| 45 | 0.0500 | 0.0550 | 0.0101 | 0.0141 | ***0.0097*** | ***0.0134*** |
| 60 | 0.0541 | 0.0589 | 0.0100 | 0.0139 | ***0.0096*** | ***0.0133*** |
| 90 | 0.0595 | 0.0641 | 0.0099 | 0.0138 | ***0.0094*** | ***0.0131*** |
| 120 | 0.0634 | 0.0676 | 0.0096 | 0.0133 | ***0.0090*** | ***0.0124*** |
| 150 | 0.0666 | 0.0705 | 0.0094 | 0.0130 | ***0.0088*** | ***0.0121*** |

*Note:* Best values are marked as ***bold italic***.

### *6.2. Long-Term Performance Decay*

Despite the promising performance of large language models in short-horizon forecasting tasks, their effectiveness deteriorates when modeling long-term dependencies. This limitation arises from a combination of factors related to data representation and architectural constraints. Given that these models are designed for NLP tasks, their training datasets likely have few long-distance dependencies and relationships (An et al., 2024). Prompt-based approaches require transforming the time series into tokenized textual sequences, often leading to structural distortions and a loss of fine-grained temporal coherence as the sequence length increases (Liu et al., 2024c). Research has found that in the presence of noise, LLMs struggle to find general signal characteristics (Bianchi et al., 2025), which is clearly our setup when considering larger and larger window sizes for financial data.

Architecturally, transformer-based LLMs suffer from fixed context windows and quadratic attention complexity, which impose practical limits on input length and introduce attention decay over distant tokens (Liu et al., 2024a; Delétang et al., 2024). These factors limit the models' ability to capture long-term dependencies, particularly in the presence of slow-moving trends, regime changes, or persistent volatility —features that are crucial in financial risk modeling. In contrast, specialized time series models leverage recursive structure, latent state variables, or hierarchical memory to maintain performance over extended horizons. Our findings thus support prior evidence that LLMs, while powerful for short-term sequence modeling, remain constrained in their ability to reason effectively over long historical windows without architectural or representational adaptations (Gruver et al., 2024; Sun et al., 2024).

Looking ahead, enhancing the robustness of LLM-based financial risk forecasting may require incorporating adaptive learning paradigms. For example, online learning frameworks — such as those proposed by Zhang et al. (2025) in dynamic localization environments or time series decomposition with LLM-Mixer (Kowsher et al., 2025) — could allow LLMs to adjust continuously to evolving market regimes.

Additionally, hybrid modeling strategies that integrate variance-constrained local–global mechanisms may help address uncertainty and heterogeneity in financial time series. Zhang et al. (2024) demonstrate the benefits of such approaches in non-stationary settings using multi-resolution modeling. Future work could explore combining the representational power of LLMs with such adaptive techniques to improve performance under volatile conditions.

## 7. Conclusions

In this paper, we introduced **LLM-VaR** and **LLM-ES**, two novel approaches for financial risk estimation using general-purpose large language models (LLMs) within the LLMTime framework. These zero-shot methods for forecasting Value at Risk (VaR) and Expected Shortfall (ES) offer a flexible and model-free alternative to traditional approaches, such as GARCH and EWMA, particularly in short-horizon, high-volatility environments.

Our empirical analysis shows that GPT-3.5 performs competitively, often outperforming both traditional econometric models and more advanced LLMs such as GPT-4 and GPT-4o in short-term VaR and ES estimation tasks. This result underscores the complex interplay between model complexity, numerical precision, and alignment with task-specific patterns. However, the performance of GPT-3.5 declines as forecast horizons increase, reflecting known limitations of Transformer-based architectures in modeling long-term dependencies (Wang et al., 2024b). In contrast, traditional models — while requiring more effort in calibration — continue to provide reliable performance for extended horizons.

Looking ahead, future work should explore the development and fine-tuning of time-series-specific LLMs that can better capture structural patterns in financial data. Promising directions include specialized models such as Chronos (Ansari et al., 2024), TimesFM (Das et al., 2024), and TimeGPT (Garza et al., 2024; Liao et al., 2024). Integrating these with established econometric frameworks like GARCH may yield more robust hybrid risk estimation systems.

In addition, recent advances in LLM alignment for time series — such as CALF (Liu et al., 2024b) —highlight promising techniques to further refine performance on numerically grounded tasks. Future work should also investigate online learning paradigms, adaptive prompt tuning time series decomposition, and the interpretability of LLM-based forecasts in regulatory settings.

Moreover, hybrid architectures that combine general-purpose LLMs with domain-specific statistical constraints, along with privacy-preserving deployment options leveraging open-weight models, represent important directions for practical adoption in regulated financial environments.

In conclusion, our findings suggest that general-purpose LLMs, particularly GPT-3.5, offer viable tools for estimating VaR and ES in contexts where agility and numerical precision are critical. While challenges remain for longer-term forecasts and interpretability, the results affirm the potential of LLMs as building blocks for next-generation financial risk analytics.

**Acknowledgements**

This paper is supported through the European Cooperation in Science & Technology COST Action grant CA19130 - Fintech and Artificial Intelligence in Finance - Towards a transparent financial industry; the project "IDA Institute of Digital Assets", CF166/15.11.2022, CN760046/23.05.2023; the project "AI for Energy Finance (AI4EFin)", CF162/15.11.2022, CN760048/23.05.2023, financed under the Romania's National Recovery and Resilience Plan, Apel nr. PNRR-III-C9-2022-I8; and the Marie Skłodowska-Curie Actions under the European Union's Horizon Europe research and innovation program for the Industrial Doctoral Network on Digital Finance, acronym DIGITAL, Project No. 101119635.

**Credit Author Statement**

Daniel Traian Pele: Conceptualization, Methodology, Supervision, Writing – original draft, review & editing. Formal analysis, Project administration. Vlad Bolovăneanu: Software, Validation, Data curation, Visualization, Methodology, Writing – review & editing. Min-Bin Lin: Formal analysis, Investigation, Writing – review & editing. Rui Ren: Data curation, Software, Methodology, Writing – review & editing. Andrei Theodor Ginavar: Software, Validation, Data curation, Visualization, Writing – review & editing. Bruno Spilak: Investigation, Validation, Writing – review & editing. Alexandru Victor Andrei: Visualization, Formal analysis, Writing – review & editing. Filip-Mihai Toma: Software, Visualization, Writing – review & editing. Stefan Lessmann: Funding acquisition, Supervision, Writing – review & editing. Wolfgang Karl Härdle: Funding acquisition, Supervision, Writing – review & editing.

**Declaration of interests**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Declaration of generative AI and AI-assisted technologies in the writing process**

During the preparation of this work, the authors used ChatGPT (OpenAI) in order to assist with language editing and clarity. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## Appendix A. Benchmark Models Specifications

The **GARCH(1,1)** model (Bollerslev, 1986) is renowned for its capacity to capture volatility clustering, a prevalent pattern in financial returns where high-volatility periods follow each other. The model is specified as follows:

$$\begin{aligned} r_t &= Z_t \sigma_t, \\ Z_t &\sim \mathcal{N}(0,1), \\ \sigma_t^2 &= \omega + \beta_1 r_{t-1}^2 + \alpha_1 \sigma_{t-1}^2, \end{aligned} \tag{A.1}$$

where $\omega > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$, and $\alpha_1 + \beta_1 < 1$. Here, $Z_t$ represents the innovation term, and $\sigma_t$ denotes the time-varying volatility, allowing the model to dynamically adjust to shifts in market volatility. For Studentized innovations, we assumed $\nu = 5$ for all models.

To further strengthen the GARCH model and improve its adaptability to sudden structural shifts, we incorporate the **Local Parametric Approach (LPA)**, which uses Local Change Point detection (Spokoiny, 1998). This approach allows the model to detect and adapt to structural breaks, enhancing its sensitivity to evolving market conditions (Spilak and Härdle, 2022).

Similarly, for the **Exponentially Weighted Moving Average** model, we go beyond the standard approach by leveraging the **Dynamic Conditional Score (DCS)** framework [7] (Creal et al., 2013; Harvey and Luati, 2014). This extended EWMA model is defined as:

$$\sigma_t^2 = (1-\lambda) u_{t-1} r_{t-1}^2 + \lambda \sigma_{t-1}^2, \tag{A.2}$$

where $u_{t-1}$ is the score term derived from the log-likelihood function, calculated as:

$$u_{t-1} = 1 + \frac{r_{t-1}^2 - \sigma_{t-1}^2}{\sigma_{t-1}^2}.$$

Under a Student's t-distribution assumption, the score term is adapted for heavy tails:

$$u_{t-1} = \frac{(\nu+1) r_{t-1}^2}{(\nu-2)\sigma_{t-1}^2 + r_{t-1}^2} - 1,$$

where $\nu$ denotes the degrees of freedom, which accounts for the heavy-tailed nature often observed in financial returns.

This advanced DCS framework is also applied to the GARCH model, allowing volatility updates to respond dynamically to shifts in the data:

$$\sigma_t^2 = \omega + \phi \sigma_{t-1}^2 + \alpha \sigma_{t-1}^2 u'_{t-1},$$

where $u'_{t-1}$ serves as a gradient term for conditional variance adjustments.

[7] Also known as the **Generalized Autoregressive Score (GAS)** framework.

## Appendix B. 1% VaR Backtesting Results

Table B1: Kupiec's POF Test for 1%VaR: p-values.

| Model | CACT | DJCI | FTSE | CRIX | CBU | SP500 | STOXX | SPGTCLTR | GDAXI |
|---|---|---|---|---|---|---|---|---|---|
| GPT-3.5.30 | 0.970 | 0.920 | 0.227 | 0.846 | 0.954 | 0.751 | 0.974 | 0.173 | 0.964 |
| GPT-3.5.45 | 0.203 | 0.077 | 0.210 | 0.379 | 0.015 | 0.001 | 0.202 | 0.064 | 0.206 |
| GPT-3.5.60 | 0.016 | 0.018 | 0.232 | 0.034 | 0.017 | 0.001 | 0.001 | 0.072 | 0.001 |
| GPT-3.5.90 | 0.001 | 0.001 | 0.102 | 0.002 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| GPT-3.5.120 | 0.002 | 0.002 | 0.029 | 0.002 | 0.002 | 0.002 | 0.001 | 0.001 | 0.002 |
| GPT-3.5.150 | 0.002 | 0.002 | 0.039 | 0.000 | 0.002 | 0.003 | 0.002 | 0.002 | 0.002 |
| GPT-4.30 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4.45 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4.60 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4.90 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4.120 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4.150 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.30 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.45 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.60 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.90 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.120 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GPT-4o.150 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GARCH.LPA | 0.517 | 0.139 | 0.030 | 0.203 | 0.274 | 0.902 | 0.520 | 0.543 | 0.157 |
| EWMA.N.80 | 0.170 | 0.032 | 0.032 | 0.000 | 0.499 | 0.782 | 0.084 | 0.183 | 0.082 |
| EWMA.N.120 | 0.964 | 0.101 | 0.103 | 0.000 | 0.209 | 0.206 | 0.426 | 0.251 | 0.230 |
| EWMA.DCS.N.80 | 0.828 | 0.095 | 0.262 | 0.292 | 0.262 | 0.020 | 0.824 | 0.798 | 0.244 |
| EWMA.DCS.N.120 | 0.611 | 0.342 | 0.128 | 0.292 | 0.338 | 0.030 | 0.608 | 0.933 | 0.315 |
| EWMA.DCS.T.80 | 0.018 | 0.020 | 0.262 | 0.292 | 0.001 | 0.001 | 0.018 | 0.001 | 0.001 |
| EWMA.DCS.T.120 | 0.001 | 0.002 | 0.128 | 0.292 | 0.002 | 0.002 | 0.001 | 0.001 | 0.001 |
| GARCH.N.120 | 0.611 | 0.383 | 0.209 | 0.001 | 0.389 | 0.101 | 0.961 | 0.128 | 0.617 |
| GARCH.N.250 | 0.919 | 0.007 | 0.864 | 0.016 | 0.097 | 0.007 | 0.923 | 0.637 | 0.910 |
| GARCH.DCS.N.120 | 0.611 | 0.206 | 0.989 | 0.001 | 0.653 | 0.383 | 0.961 | 0.060 | 0.617 |
| GARCH.DCS.N.250 | 0.919 | 0.007 | 0.864 | 0.033 | 0.097 | 0.007 | 0.923 | 0.637 | 0.910 |
| GARCH.DCS.T.120 | 0.311 | 0.002 | 0.338 | 0.506 | 0.128 | 0.002 | 0.001 | 0.108 | 0.118 |
| GARCH.DCS.T.250 | 0.308 | 0.007 | 0.097 | 0.506 | 0.007 | 0.007 | 0.006 | 0.005 | 0.006 |

*Note:* 1. Green: p-values higher than 0.05. 2. Red: p-values lower than 0.05.

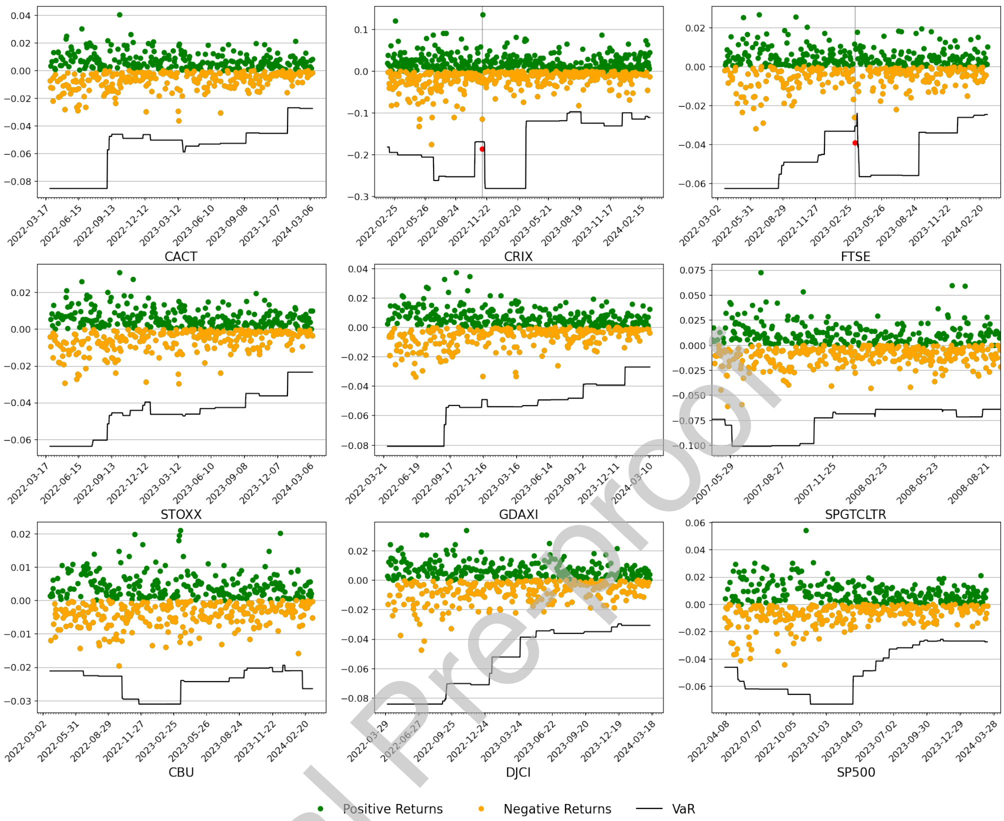



Figure B1: VaR Exceedances for LLM-VaR GPT-3.5 120-day rolling window.

*Note:* Color codes: Red dots indicate model failures, green dots show gains, and orange dots show losses within the expected range.

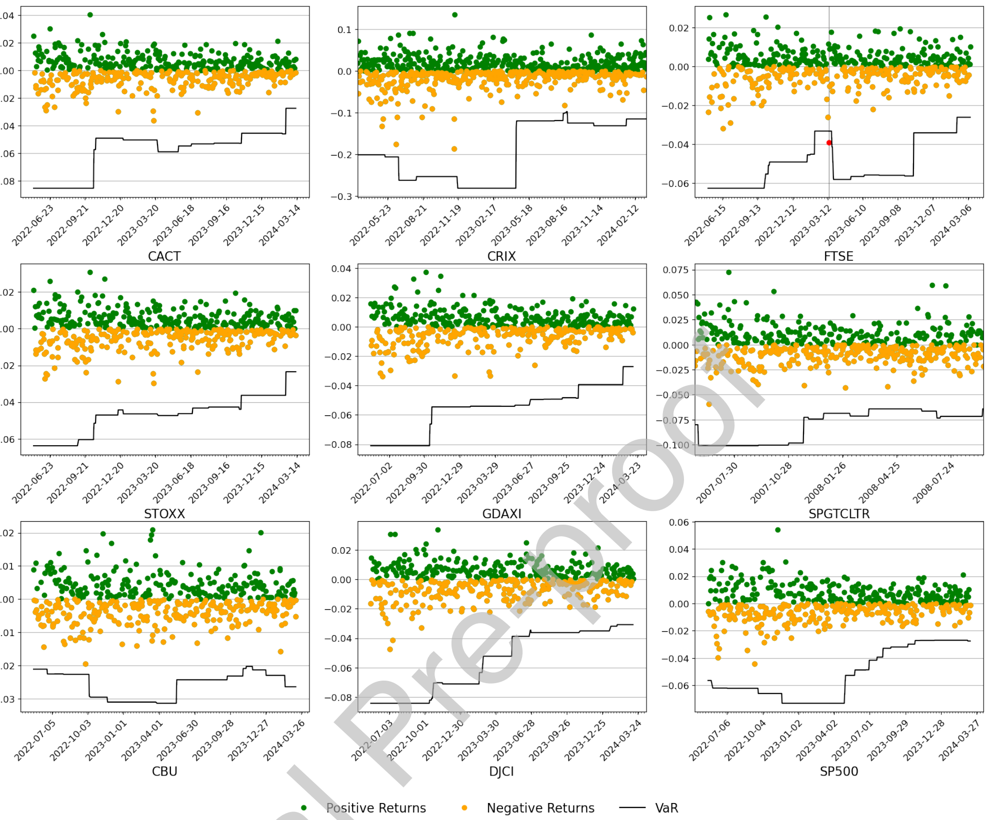



Figure B2: VaR Exceedances for LLM-VaR GPT-3.5 150-day rolling window.

*Note:* Color codes: Red dots indicate model failures, green dots show gains, and orange dots show losses within the expected range.

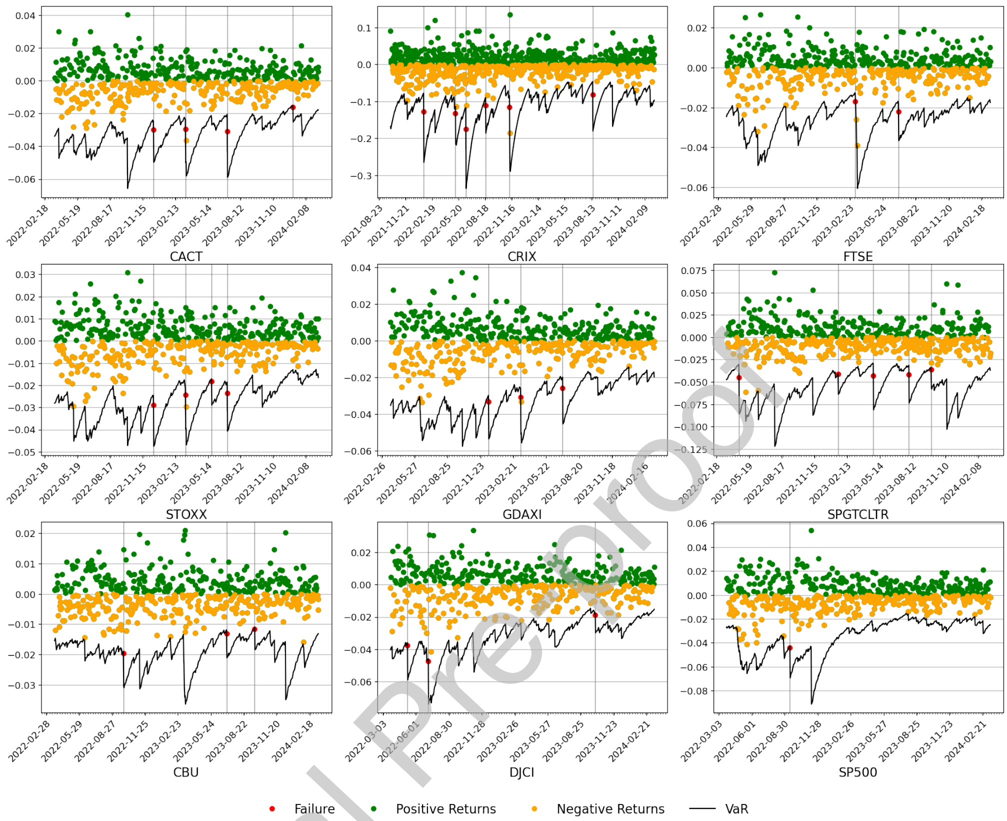



Figure B3: VaR Exceedances for EWMA-DCS (120-day normal innovations).

*Note:* Color codes: Red dots indicate model failures, green dots show gains, and orange dots show losses within the expected range.

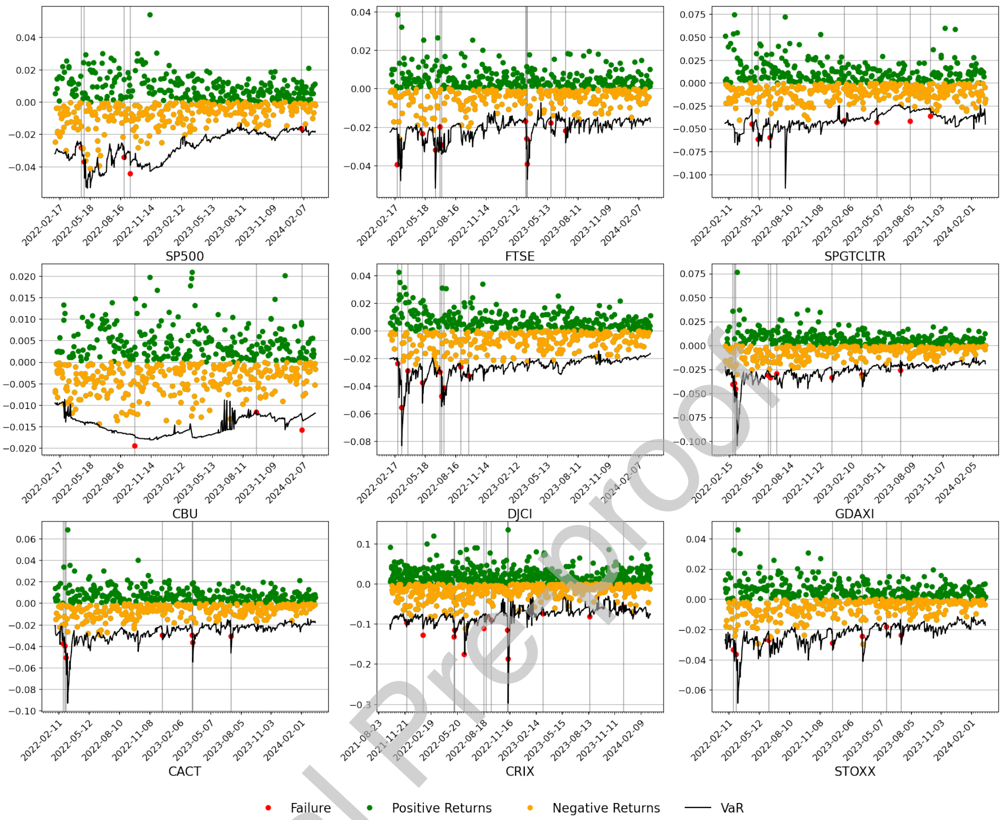



Figure B4: VaR Exceedances for GARCH-LPA.

*Note:* Color codes: Red dots indicate model failures, green dots show gains, and orange dots show losses within the expected range.

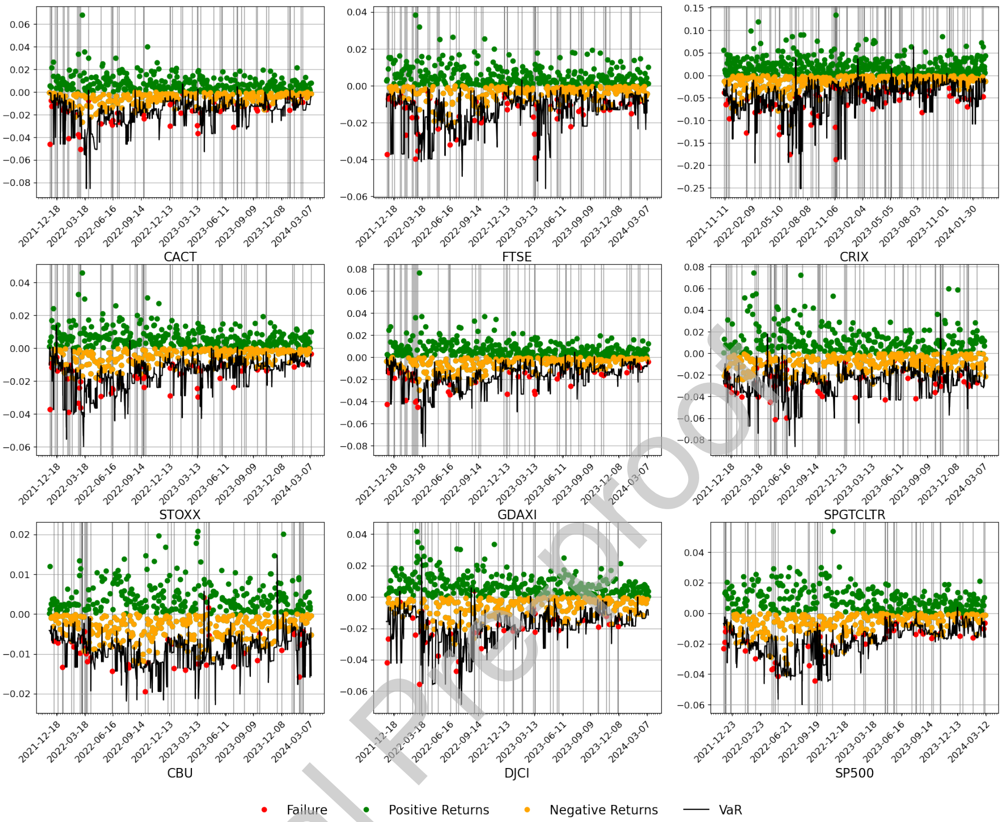



Figure B5: VaR Exceedances for LLM-VaR GPT-4 30-day rolling window.

*Note:* Color codes: Red dots indicate model failures, green dots show gains, and orange dots show losses within the expected range.

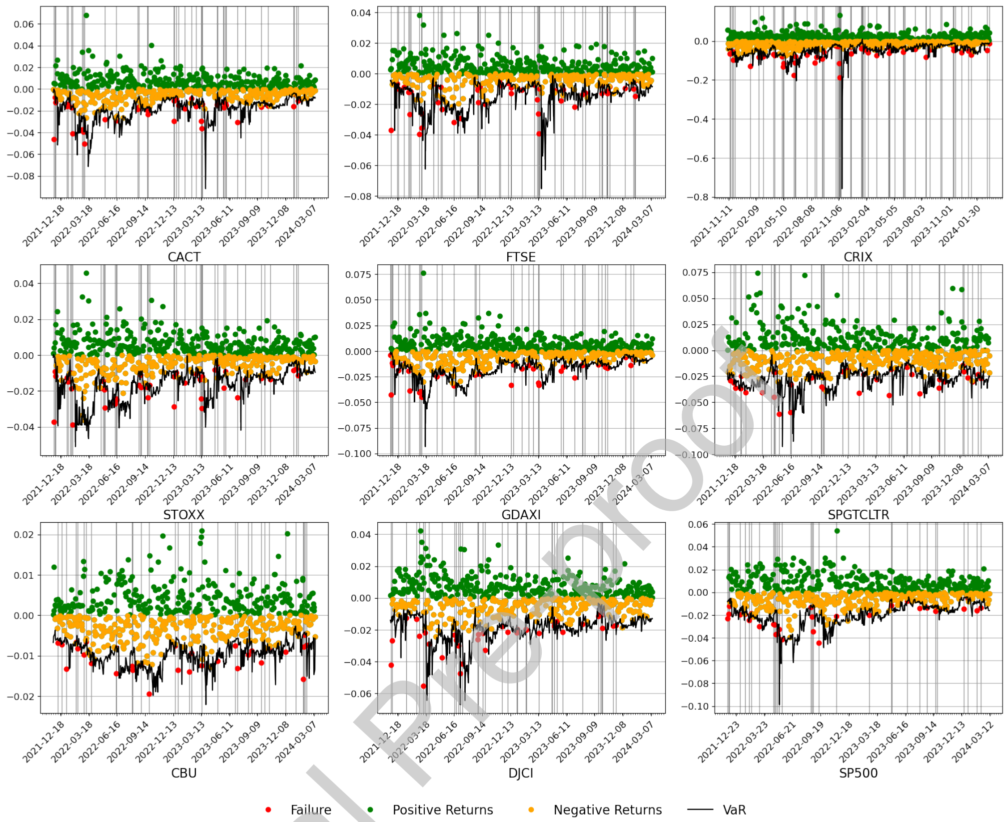



Figure B6: VaR Exceedances for LLM-VaR GPT-4o 30-day rolling window.

*Note:* Color codes: Red dots indicate model failures, green dots show gains, and orange dots show losses within the expected range.

## Appendix C. LLM Distribution plots

We present Kernel Density Estimation plots for the empirical log return distributions and LLM predictions for each index. We chose the 30-day window for illustration purposes. For higher windows, we observe that distributions become tighter (values close to 0 are predicted more frequently). This explains general poor results for larger windows from an empirical angle.

GPT-4 and GPT-4o show a better fit of the distribution overall, but, as we saw when backtesting, VaR and ES predictions are better calibrated for GPT-3.5.

]

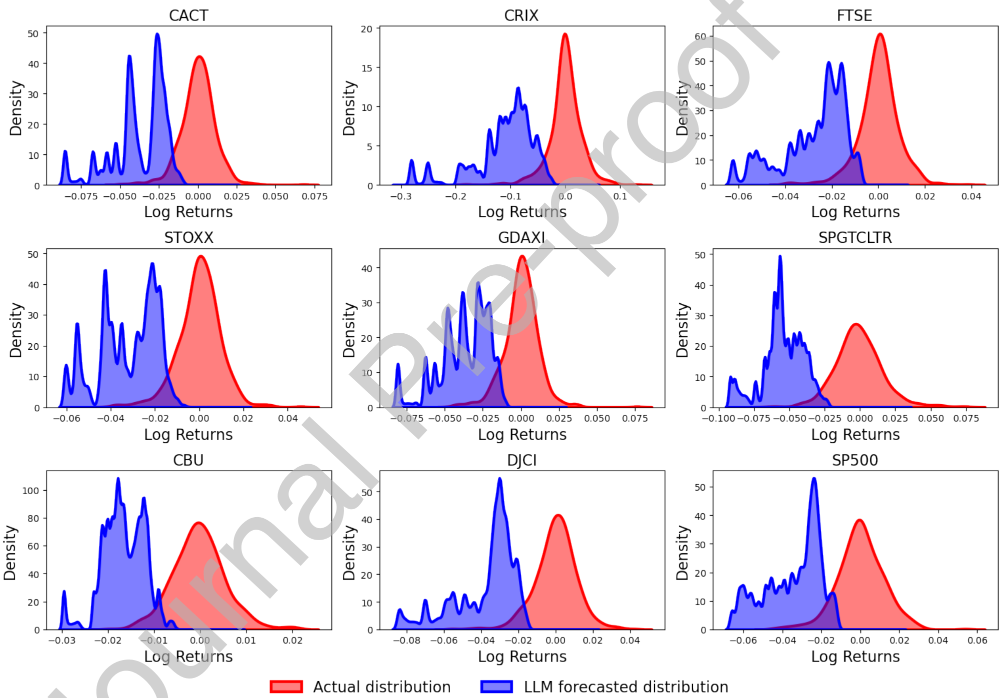



Figure C1: KDE Estimation for GPT-3.5 30-day rolling window predictions and actual log returns.

*Note:* Color codes: Red indicates log return distribution, blue show LLM predictions

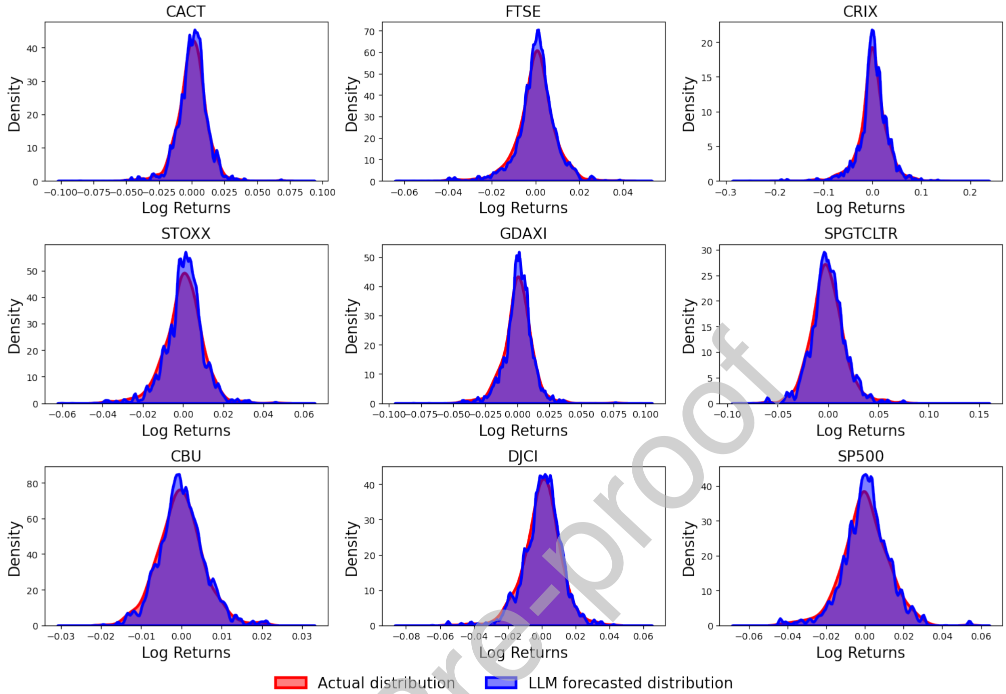



Figure C2: KDE Estimation for GPT-4 30-day rolling window predictions are actual log returns.

*Note:* Color codes: Blue indicates log return distribution, red show LLM predictions

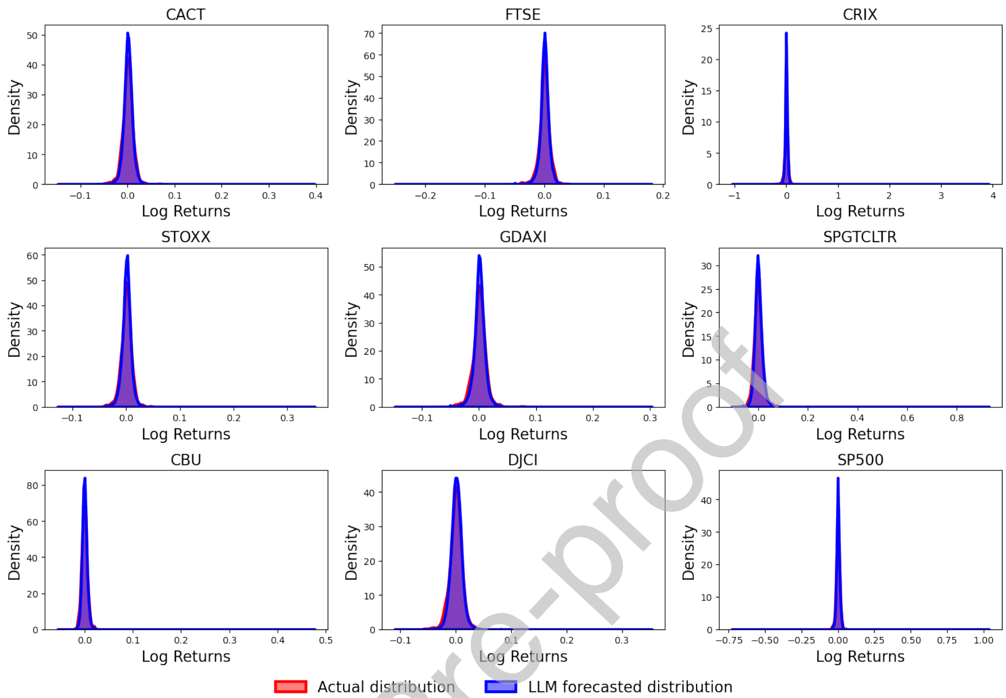



Figure C3: KDE Estimation for GPT-4o 30-day rolling window predictions are actual log returns.

*Note:* Color codes: Blue indicates log return distribution, red show LLM predictions

## References


Acerbi, C., Székely, B., 2014. Back-testing expected shortfall. Risk , 76–81URL: https://www.proquest.com/scholarly-journals/back-testing-expected-shortfall/docview/1636546980/se-2.

Ahmed, S., Nielsen, I.E., Tripathi, A., Siddiqui, S., Ramachandran, R.P., Rasool, G., 2023. Transformers in Time-Series Analysis: A Tutorial. Circuits, Systems, and Signal Processing 42, 7433–7466. URL: https://doi.org/10.1007/s00034-023-02454-8, doi:10.1007/s00034-023-02454-8.

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, 393–427. URL: https://doi.org/10.1080/14697688.2022.2159505, doi:10.1080/14697688.2022.2159505, arXiv:https://doi.org/10.1080/14697688.2022.2159505.

An, C., Zhang, J., Zhong, M., Li, L., Gong, S., Luo, Y., Xu, J., Kong, L., 2024. Why does the effective context length of llms fall short? URL: https://arxiv.org/abs/2410.18745, arXiv:2410.18745.

Ansari, A.F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S.S., Arango, S.P., Kapoor, S., Zschiegner, J., Maddix, D.C., Wang, H., Mahoney, M.W., Torkkola, K., Wilson, A.G., Bohlke-Schneider, M., Wang, Y., 2024. Chronos: Learning the Language of Time Series. URL: https://arxiv.org/abs/2403.07815, arXiv:2403.07815.

Basel Committee on Banking Supervision, 1996. Supervisory framework for the use of backtesting in conjunction with the internal models approach to market risk capital requirements. Technical Report. Bank for International Settlements. URL: https://www.bis.org/publ/bcbs22.htm.

Bento, J., Saleiro, P., Cruz, A.F., Figueiredo, M.A.T., Bizarro, P., 2021. Timeshap: Explaining recurrent models through sequence perturbations, in: Zhu, F., Ooi, B.C., Miao, C. (Eds.), KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021, ACM. pp. 2565–2573. URL: https://doi.org/10.1145/3447548.3467166, doi:10.1145/3447548.3467166.

Bhatia, G., Nagoudi, E.M.B., Cavusoglu, H., Abdul-Mageed, M., 2024. FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models, in: Annual Meeting of the Association for Computational Linguistics. URL: https://api.semanticscholar.org/CorpusID:267750823.

Bianchi, O., Koretsky, M.J., Willey, M., Alvarado, C.X., Nayak, T., Asija, A., Kuznetsov, N., Nalls, M.A., Faghri, F., Khashabi, D., 2025. Lost in the haystack: Smaller needles are more difficult for llms to find. URL: https://arxiv.org/abs/2505.18148, arXiv:2505.18148.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327. URL: https://www.sciencedirect.com/science/article/pii/0304407686900631, doi:https://doi.org/10.1016/0304-4076(86)90063-1.

Cao, R., Wang, Q., 2024. An evaluation of standard statistical models and llms on time series forecasting. URL: https://arxiv.org/abs/2408.04867, arXiv:2408.04867.

Cao, Y., Chen, Z., Kumar, P., Pei, Q., Yu, Y., Li, H., Dimino, F., Ausiello, L., Subbalakshmi, K.P., Ndiaye, P.M., 2025. RiskLabs: Predicting Financial Risk Using Large Language Model based on Multimodal and Multi-Sources Data. URL: https://arxiv.org/abs/2404.07452, arXiv:2404.07452.

Christoffersen, P.F., 1998. Evaluating Interval Forecasts. International Economic Review 39, 841–862. URL: https://ideas.repec.org/a/ier/iecrev/v39y1998i4p841-62.html.

Clift, S.S., Costanzino, N., Curran, M., 2016. Empirical Performance of Backtesting Methods for Expected Shortfall. Econometrics: Econometric & Statistical Methods - Special Topics eJournal URL: https://api.semanticscholar.org/CorpusID:155198492.

Costanzino, N., Curran, M., 2015. Backtesting General Spectral Risk Measures with Application to Expected Shortfall. URL: https://papers.ssrn.com/abstract=2514403, doi:10.2139/ssrn.2514403.

Costanzino, N., Curran, M., 2018. A Traffic Light Approach to Back-Testing Expected Shortfall. Risk.net Available at: https://www.risk.net/.

Creal, D., Koopman, S.J., Lucas, A., 2013. GENERALIZED AUTOREGRESSIVE SCORE MODELS WITH APPLICATIONS. Journal of Applied Econometrics 28, 777–795. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.1279, doi:https://doi.org/10.1002/jae.1279, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.1279.

Das, A., Kong, W., Sen, R., Zhou, Y., 2024. A decoder-only foundation model for time-series forecasting. URL: https://arxiv.org/abs/2310.10688, arXiv:2310.10688.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J.L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R.J., Jin, R.L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S.S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W.L., Zeng, W., Zhao, W., An, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Li, X.Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Zhang, X., Chen, X., Nie, X., Sun, X., Wang, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Song, X., Shan, X., Zhou, X., Yang, X., Li, X., Su, X., Lin, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhu, Y.X., Zhang, Y., Xu, Y., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Yu, Y., Zheng, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Tang, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Wu, Y., Ou, Y., Zhu, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Zha, Y., Xiong, Y., Ma, Y., Yan, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z.F., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Huang, Z., Zhang, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Xu, Z., Wu, Z., Zhang, Z., Li, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Gao, Z., Pan, Z., 2025. Deepseek-v3 technical report. URL: https://arxiv.org/abs/2412.19437, arXiv:2412.19437.

Delétang, G., Ruoss, A., Duquenne, P.A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L.K., Aitchison, M., Orseau, L., Hutter, M., Veness, J., 2024. Language modeling is compression. URL: https://arxiv.org/abs/2309.10668, arXiv:2309.10668.

Fatouros, D., Sermpinis, G., Stasinakis, C., Dunis, C., 2023. DeepVaR: A deep learning approach for Value at Risk forecasting. Expert Systems with Applications 213, 119104. URL: https://doi.org/10.1016/j.eswa.2022.119104, doi:10.1016/j.eswa.2022.119104.

Garza, A., Challu, C., Mergenthaler-Canseco, M., 2024. TimeGPT-1 URL: https://arxiv.org/abs/2310.03589, arXiv:2310.03589.

Goel, A., Pasricha, P., Kanniainen, J., 2025a. Time-Series Foundation AI Model for Value-at-Risk Forecasting. URL: https://arxiv.org/abs/2410.11773, arXiv:2410.11773.

Goel, A., Pasricha, P., Magris, M., Kanniainen, J., 2025b. Foundation Time-Series AI Model for Realized Volatility Forecasting. URL: https://arxiv.org/abs/2505.11163, arXiv:2505.11163.

Google DeepMind, 2024. Gemma 2: Lightweight, Open Models from Google DeepMind. https://deepmind.google/technologies/gemma. Accessed: 25-May-2025.

Gruver, N., 2025. LLMTime Github Repository. https://github.com/ngruver/llmtime. Accessed: 05-Jun-2025.

Gruver, N., Finzi, M., Qiu, S., Wilson, A.G., 2024. Large language models are zero-shot time series forecasters, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. URL: https://dl.acm.org/doi/10.5555/3666122.3666983.

Harvey, A., Luati, A., 2014. Filtering With Heavy Tails. Journal of the American Statistical Association 109, 1112–1122. URL: https://EconPapers.repec.org/RePEc:taf:jnlasa:v:109:y:2014:i:507:p:1112-1122.

Kowsher, M., Sobuj, M.S.I., Prottasha, N.J., Alanis, E.A., Garibay, O.O., Yousefi, N., 2025. Llm-mixer: Multiscale mixing in llms for time series forecasting. URL: https://arxiv.org/abs/2410.11674, arXiv:2410.11674.

Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives 3, 73–84. doi:https://doi.org/10.3905/jod.1995.407942.

Lagasio, V., Pirillo, J., Belloli, M., 2025. Integrating generative ai and large language models in financial sector risk management: Regulatory frameworks and practical applications. Risk Management Magazine doi:10.47473/2020rmm0150.

Lazar, E., Zhang, N., 2019. Model risk of expected shortfall. Journal of Banking & Finance 105, 74–93. URL: https://EconPapers.repec.org/RePEc:eee:jbfina:v:105:y:2019:i:c:p:74-93.

Lee, H., 2025. Unleashing the potential of large language models in the finance industry doi:10.31219/osf.io/ahkd3.

Li, W., Liu, W., Deng, M., Liu, X., Feng, L., 2025. The impact of large language models on accounting and future application scenarios. Journal of Accounting Literature doi:10.1108/jal-12-2024-0357.

Li, Y., Wang, S., Ding, H., Chen, H., 2023. Large Language Models in Finance: A Survey. arXiv preprint arXiv:2311.10723 URL: http://arxiv.org/pdf/2311.10723.

Liao, W., Porte-Agel, F., Fang, J., Rehtanz, C., Wang, S., Yang, D., Yang, Z., 2024. TimeGPT in Load Forecasting: A Large Time Series Model Perspective. URL: https://arxiv.org/abs/2404.04885, arXiv:2404.04885.

Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P., 2024a. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics 12, 157–173. URL: https://aclanthology.org/2024.tacl-1.9/, doi:10.1162/tacl_a_00638.

Liu, P., Guo, H., Dai, T., Li, N., Bao, J., Ren, X., Jiang, Y., Xia, S.T., 2024b. CALF: Aligning LLMs for Time Series Forecasting via Cross-modal Fine-Tuning. URL: https://arxiv.org/abs/2403.07300, arXiv:2403.07300.

Liu, Y., 2025. The development of large language models in the financial field. Proceedings of Business and Economic Studies doi:10.26689/pbes.v8i2.10267.

Liu, Y., Qin, G., Huang, X., Wang, J., Long, M., 2024c. Autotimes: Autoregressive time series forecasters via large language models. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105000527493&partnerID=40&md5=258cd9b796d7ffa0a0fd21f00dd42ce0.

McCaul, E., 2024. From data to decisions: AI and supervision. Article for *Revue Banque*. URL: https://www.bankingsupervision.europa.eu/press/interviews/date/2024/html/ssm.in240226~c6f7fc9251.en.html. Elizabeth McCaul is a member of the Supervisory Board of the ECB.

McNeil, A.J., Frey, R., Embrechts, P., 2005. Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press.

Mistral AI, 2024. Mistral Large 2: Open-weight High-Performance Language Model. https://mistral.ai/news/mistral-large/. Accessed: 25-May-2025.

Nie, Y., Kong, Y., Dong, X., Mulvey, J.M., Poor, H.V., Wen, Q., Zohren, S., 2024. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. URL: https://arxiv.org/abs/2406.11903, arXiv:2406.11903.

OpenAI, 2023. GPT-4 Technical Report. OpenAI Technical Report Available at https://openai.com/research/gpt-4.

OpenAI, 2024. GPT-3.5 Turbo Technical Report. Available at https://www.openai.com/research/gpt-3-5.

OpenAI, 2025a. OpenAI API Privacy. https://openai.com/enterprise-privacy/. Accesse: 26-May-2025.

OpenAI, 2025b. OpenAI Chat API. https://platform.openai.com/docs/api-reference/responses/create#responses-create-temperature. Accessed: 05-Jun-2025.

OpenAI, 2025c. OpenAI Model Tokenizer. https://platform.openai.com/tokenizer. Accessed: 05-Jun-2025.

Qiu, Z., Lazar, E., Nakata, K., 2024. VaR and ES forecasting via recurrent neural network-based stateful models. International Review of Financial Analysis 92, 103102. URL: https://www.sciencedirect.com/science/article/pii/S1057521924000346, doi:https://doi.org/10.1016/j.irfa.2024.103102.

Spilak, B., Härdle, W.K., 2022. Tail-Risk Protection: Machine Learning Meets Modern Econometrics, in: Lee, C.F., Lee, A.C. (Eds.), Encyclopedia of Finance. Springer. Springer Books. chapter 92, pp. 2177–2211. URL: https://ideas.repec.org/h/spr/sprchp/978-3-030-91231-4_94.html, doi:10.1007/978-3-030-91231-4.

Spokoiny, V.G., 1998. Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. The Annals of Statistics 26, 1356 – 1378. URL: https://doi.org/10.1214/aos/1024691246, doi:10.1214/aos/1024691246.

Sun, C., Li, H., Li, Y., Hong, S., 2024. TEST: Text Prototype Aligned Embedding to Activate LLM's Ability for Time Series, in: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=Tuh4nZVb0g.

Tang, H., Zhang, C., Jin, M., Yu, Q., Wang, Z., Jin, X., Zhang, Y., Du, M., 2025. Time Series Forecasting with LLMs: Understanding and Enhancing Model Capabilities. SIGKDD Explor. Newsl. 26, 109–118. URL: https://doi.org/10.1145/3715073.3715083, doi:10.1145/3715073.3715083.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., et al., 2024. LLaMA 3: Open Foundation and Instruction Models. https://ai.meta.com/llama. Meta AI, Accessed: 25-May-2025.

Trachova, D., Lysak, O., 2025. Large language models in financial statement analysis: A systematic review of recent advances, practical implications, and future research. Scientific papers OF DMYTRO MOTORNYI TAVRIA STATE AGROTECHNOLOGICAL UNIVERSITY (ECONOMIC SCIENCES) , 40–46doi:10.32782/2519-884X-2025-54-5.

Wang, J., Wang, S., Lv, M., Yang, X., Fang, Y., 2024a. Forecasting VaR and ES by using deep quantile regression, GANs-based scenario generation, and heterogeneous market hypothesis. Financial Innovation 10, 36. URL: https://doi.org/10.1186/s40854-023-00564-5, doi:10.1186/s40854-023-00564-5.

Wang, X., Salmani, M., Omidi, P., Ren, X., Rezagholizadeh, M., Eshaghi, A., 2024b. Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models. URL: https://arxiv.org/abs/2402.02244, arXiv:2402.02244.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z., 2025. Qwen3 technical report. URL: https://arxiv.org/abs/2505.09388, arXiv:2505.09388.

Zhang, J., Li, Y., Li, Q., Xiao, W., 2024. Variance-Constrained Local–Global Modeling for Device-Free Localization Under Uncertainties. IEEE Transactions on Industrial Informatics 20, 5229–5240. doi:10.1109/TII.2023.3281234.

Zhang, J., Xue, J., Li, Y., Cotton, S.L., 2025. Leveraging Online Learning for Domain-Adaptation in Wi-Fi-Based Device-Free Localization. IEEE Transactions on Mobile Computing doi:10.1109/TMC.2025.3552538. advance online publication.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., Jin, R., 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. URL: https://arxiv.org/abs/2201.12740, arXiv:2201.12740.