

# **FTSE 100 Index Returns**

## **Project Report**

### **Submitted by:**

Geet Singhi (22b1035)

Jay Mehta (22b1281)

Piyush Babar (22b3909)

Ved Danait (22b1818)

### **Course:**

EC602 - Time Series Econometrics for Economic Analysis - I

### **Institution:**

Indian Institute of Technology, Bombay

November 15, 2025

# Contents

<b>1</b>	<b>Introduction and Objectives</b>	<b>2</b>
<b>2</b>	<b>Methodology and Data</b>	<b>3</b>
2.1	Dataset Description . . . . .	3
2.1.1	Data Preprocessing . . . . .	3
2.2	Exploratory Analysis . . . . .	3
2.2.1	Summary Statistics . . . . .	3
2.2.2	Distribution and Normality . . . . .	4
2.2.3	Volatility Clustering and Serial Dependence . . . . .	5
2.2.4	Stationarity and Heteroskedasticity Tests . . . . .	5
2.3	Modeling Framework . . . . .	5
2.3.1	Mean Equation: ARMA Specification . . . . .	5
2.3.2	Variance Equation: Conditional Heteroskedasticity Models . . . . .	6
2.3.3	Distributional Assumptions . . . . .	7
2.3.4	Modeling Strategy . . . . .	7
<b>3</b>	<b>Results and Analysis</b>	<b>8</b>
3.1	Model Estimation . . . . .	8
3.1.1	Model Selection Criteria . . . . .	8
3.1.2	Basic Model Results . . . . .	8
3.1.3	Advanced Model Results . . . . .	9
3.1.4	Key Insights . . . . .	10
3.2	Return Forecasting . . . . .	10
3.2.1	Model Fitting Results . . . . .	10
3.2.2	Point Forecasts: Return and Volatility . . . . .	11
3.2.3	Multi-Horizon Volatility Forecasts . . . . .	12
3.2.4	Discussion . . . . .	13
3.3	Out-of-Sample Testing and Backtesting . . . . .	14
3.3.1	Forecast Accuracy Metrics . . . . .	14
3.3.2	Model Comparison Results . . . . .	15
3.3.3	Analysis of Model Features . . . . .	17
3.3.4	Model Ranking and Recommendations . . . . .	17
<b>4</b>	<b>Conclusion</b>	<b>18</b>
4.1	Summary of Main Results . . . . .	18
4.1.1	Stylized Facts and Model Justification . . . . .	18
4.1.2	In-Sample Model Selection . . . . .	18
4.1.3	Out-of-Sample Forecasting Performance . . . . .	18
4.1.4	Key Finding: Overfitting vs Generalization . . . . .	19

# 1. Introduction and Objectives

Financial time series, such as stock market returns, often display distinct patterns that differ from typical datasets. For instance, large movements in prices tend to cluster together over time (a phenomenon known as volatility clustering), and the distribution of returns often deviates from the normal distribution, showing heavier tails and asymmetry.

The objective of this project is to model and forecast the daily returns of the FTSE 100 index for the period 2005–2007. By doing so, we aim to capture both the average behaviour of returns and the way volatility changes over time, under different model and distributional assumptions. Understanding these dynamics is essential for risk management, portfolio design, and forecasting future market behaviour.

The specific objectives of the project are as follows:

- To clean and prepare the FTSE 100 data for analysis.
- To perform exploratory data analysis (EDA) to understand return characteristics and determine suitable modelling approaches.
- To model the conditional mean of returns using an appropriate time-series model.
- To model the conditional variance (volatility) using volatility models such as GARCH and its variants.
- To evaluate model performance using statistical information criteria and diagnostic tests.
- To forecast volatility and interpret the persistence of risk in financial markets.

## 2. Methodology and Data

### 2.1 Dataset Description

The dataset used in this project consists of daily closing prices of the FTSE 100 index, obtained from the London Stock Exchange for the period 2005–2007. The FTSE 100 is a stock market index that measures the performance of the 100 largest companies listed on the London Stock Exchange by market capitalization. The data captures day-to-day movements in the index, providing a basis for modelling returns and volatility dynamics.

#### 2.1.1 Data Preprocessing

Preprocessing was implemented in Python using *pandas* and *numpy*. The raw dataset contained several columns such as *Open*, *High*, *Low*, *Change %*, and *Vol.*, which were dropped to retain only the *Date* and *Price* columns. We then converted dates to a standard datetime format, sorted them chronologically, and cast prices to numeric values. Missing or invalid entries were dropped.

Daily simple returns were computed as

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}},$$

where  $P_t$  is the closing price on day  $t$ . Log-returns were then defined as

$$\ell_t = \ln(1 + r_t).$$

### 2.2 Exploratory Analysis

#### 2.2.1 Summary Statistics

Summary statistics of daily returns and log-returns for the FTSE 100 index (2005–2007) are shown in Table 2.1. Both series have a mean close to zero and similar standard deviations (approximately 0.84%). Skewness is slightly negative, indicating a longer left tail, while the kurtosis values exceed 5, suggesting leptokurtic (fat-tailed) behavior.

Table 2.1: Descriptive Statistics of FTSE 100 Daily Returns (2005–2007)

Series	Mean	Std. Dev.	Skewness	Kurtosis
Returns	0.000415	0.00843	-0.373	5.711
Log Returns	0.000379	0.00844	-0.431	5.789

## 2.2.2 Distribution and Normality

Histograms and kernel density estimates for returns and log-returns (Figure 2.1) are centered near zero but display visibly heavier tails than a Gaussian. Q–Q plots against the normal distribution show clear tail deviations, whereas the fit against a Student- $t$  distribution with about five degrees of freedom is substantially better, especially in the extremes (Figures 2.2 and 2.3).

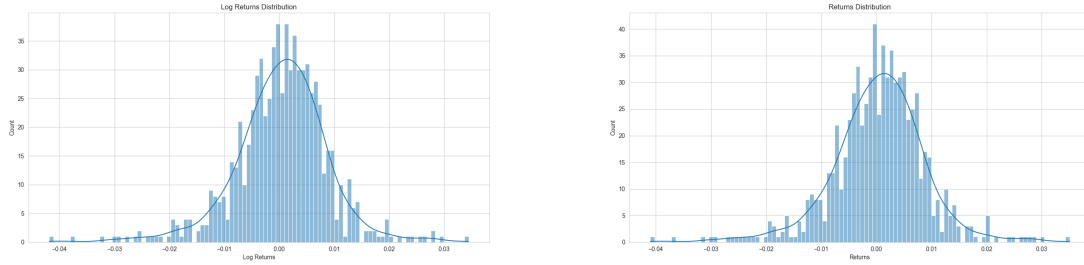


Figure 2.1: Histogram and KDE plots for Returns (left) and Log Returns (right).

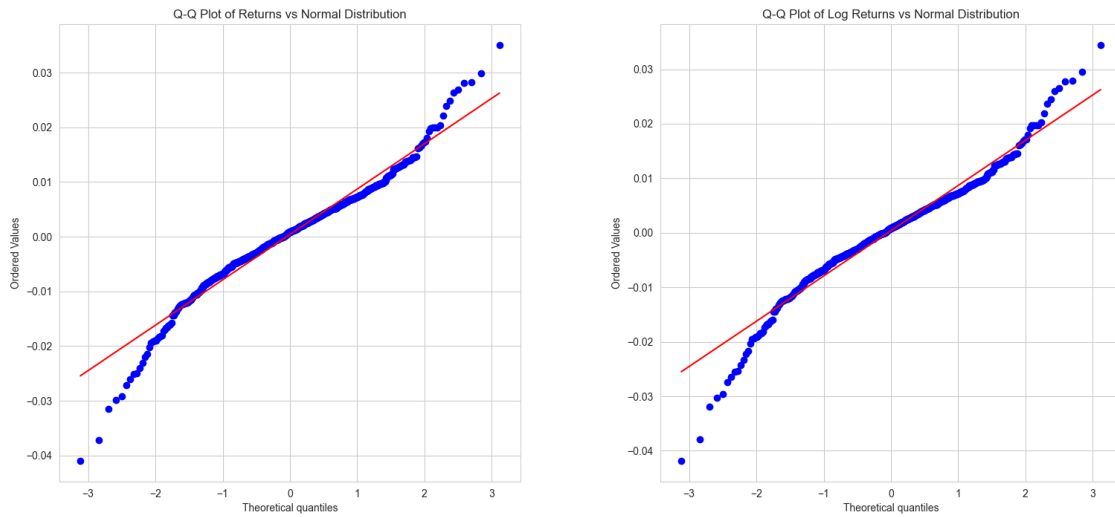


Figure 2.2: Q–Q plots of Returns and Log Returns vs Normal Distribution.

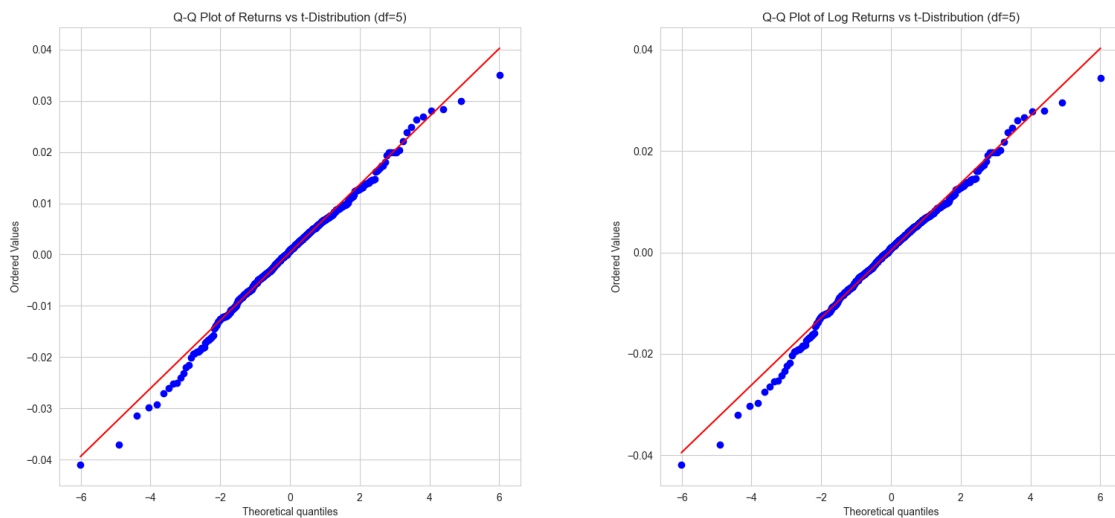


Figure 2.3: Q–Q plots of Returns and Log Returns vs Student- $t$  Distribution.

### 2.2.3 Volatility Clustering and Serial Dependence

Volatility clustering is a key stylized fact in financial time series—large price changes tend to be followed by large changes (of either sign). While ACF and PACF plots of returns (Figure 2.4) show very marginal autocorrelation, the squared returns exhibit a slowly decaying ACF, indicating persistent volatility dynamics.

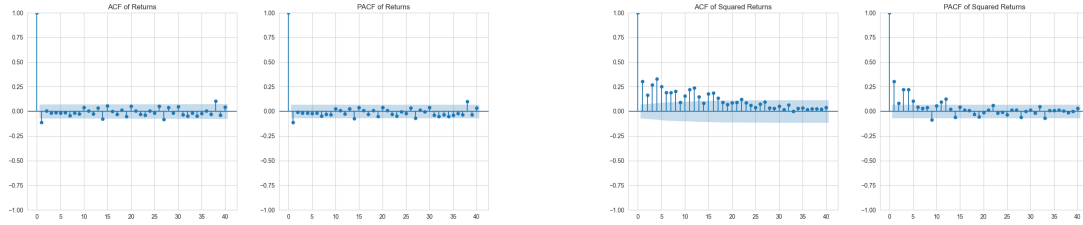


Figure 2.4: ACF/PACF of Returns (left) and Squared Returns (right).

### 2.2.4 Stationarity and Heteroskedasticity Tests

Formal tests were used to confirm these visual findings:

- **ADF test:** Price series is non-stationary ( $p > 0.05$ ); returns and log-returns are strongly stationary ( $p < 0.01$ ).
- **ARCH-LM test:** Significant ARCH effects detected ( $p < 0.001$ ), confirming volatility clustering.
- **Ljung-Box test:** Returns show weak but statistically significant autocorrelation at higher lags ( $p \approx 0.04$ ).

These results imply that the conditional mean of returns exhibits little dependence, while the conditional variance is serially correlated. This justifies modeling volatility explicitly using ARCH/GARCH-type models.

Overall, the exploratory analysis supports the use of an ARMA–GARCH framework for the FTSE 100 return series, with Student- $t$  innovations to better capture tail behavior.

## 2.3 Modeling Framework

Financial return series often exhibit weak serial correlation in the mean but strong persistence in volatility (heteroskedasticity). To capture these dynamics, we adopt a two-step modeling framework that jointly estimates the conditional mean and variance of the returns process.

### 2.3.1 Mean Equation: ARMA Specification

The conditional mean of daily returns is modeled using an autoregressive moving average process:

$$r_t = \mu + \phi_1 r_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t,$$

where  $r_t$  denotes the return at time  $t$ ,  $\mu$  is the unconditional mean,  $\varepsilon_t$  is the innovation term (white noise),  $\phi_1$  captures the autoregressive effect, and  $\theta_1$  represents the moving average effect. This ARMA(1,1) specification allows the conditional mean to account for short-term serial dependence in returns while leaving the volatility dynamics to the variance equation.

### 2.3.2 Variance Equation: Conditional Heteroskedasticity Models

The conditional variance  $\sigma_t^2 = \text{Var}(\varepsilon_t | \mathcal{F}_{t-1})$  is modeled using various members of the GARCH family. We first estimate the standard GARCH model and then extend it with nonlinear and asymmetric variants.

#### GARCH(1,1)

The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model proposed by Bollerslev (1986) is specified as:

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

where:

- $\omega > 0$  is a constant term,
- $\alpha_1$  measures the short-run reaction of volatility to market shocks (the ARCH effect),
- $\beta_1$  captures the persistence of volatility (the GARCH effect).

High values of  $\alpha_1 + \beta_1$  close to one imply persistent volatility clustering, a common feature of financial time series.

#### ARCH(1)

As a baseline, we also estimate the simpler ARCH(1) model (Engle, 1982):

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2,$$

which models the conditional variance purely as a function of past squared residuals. This captures short-run volatility but typically underestimates long-term persistence.

#### EGARCH(1,1)

The Exponential GARCH model (Nelson, 1991) models the logarithm of conditional variance:

$$\ln(\sigma_t^2) = \omega + \beta_1 \ln(\sigma_{t-1}^2) + \alpha_1 \left| \frac{\varepsilon_{t-1}}{\sigma_{t-1}} \right| + \gamma_1 \frac{\varepsilon_{t-1}}{\sigma_{t-1}}.$$

This formulation ensures positivity of variance without parameter restrictions. The term  $\gamma_1$  introduces asymmetry: negative shocks (bad news) and positive shocks (good news) can have different effects on volatility, consistent with the “leverage effect” observed in equity returns.

#### GJR–GARCH(1,1)

The Glosten–Jagannathan–Runkle GARCH model (GJR–GARCH; 1993) introduces asymmetry through an indicator variable:

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 I_{\{\varepsilon_{t-1} < 0\}} \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

where  $I_{\{\varepsilon_{t-1} < 0\}}$  equals 1 if the previous shock was negative and 0 otherwise. The additional term  $\gamma_1$  captures the differential impact of negative shocks on volatility. If  $\gamma_1 > 0$ , then negative shocks increase volatility more than positive shocks of the same magnitude—another representation of the leverage effect.

### 2.3.3 Distributional Assumptions

The innovation terms  $\varepsilon_t$  represent the unpredictable shocks in returns and are modeled under different distributional assumptions to capture the empirical characteristics of financial data.

- **Normal distribution:**  $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ . This is the standard assumption, implying symmetric and thin-tailed innovations.
- **Student- $t$  distribution:**  $\varepsilon_t \sim t_\nu(0, \sigma_t^2)$ . The Student- $t$  allows for heavier tails through its degrees of freedom parameter  $\nu$ , making it more suitable for capturing large shocks and volatility bursts often seen in financial returns.
- **Skewed Student- $t$  distribution:**  $\varepsilon_t \sim \text{Skew-}t(\nu, \xi)$ . This distribution introduces an additional skewness parameter  $\xi$  to allow for asymmetric behavior in returns. Empirically, negative returns in financial markets often occur with higher magnitude and frequency than positive ones; the skew- $t$  specification accommodates this asymmetry by adjusting the shape of the distribution.

Overall, the Student- $t$  and Skewed- $t$  specifications provide greater flexibility in modeling fat-tailed and asymmetric return distributions, which are both key features of real-world financial data.

### 2.3.4 Modeling Strategy

Two modeling stages were implemented:

1. **Basic Models:** ARMA(1,1)–GARCH(1,1) combinations were estimated under both Normal and Student- $t$  error distributions to capture conditional mean and volatility.
2. **Advanced Models:** In the advanced stage, we estimated a range of volatility models including ARCH(1), GARCH(1,1), EGARCH(1,1), and GJR–GARCH(1,1), along with selected higher-order GARCH( $p, q$ ) variants. These models were fitted under both Student- $t$  and Skewed Student- $t$  distributions to account for heavy tails, asymmetry, and potential nonlinearities in the volatility process.

## 3. Results and Analysis

### 3.1 Model Estimation

Model estimation was conducted using Python’s `arch` and `statsmodels` libraries. A range of ARMA – GARCH family models were estimated to capture both the conditional mean and variance dynamics of FTSE 100 returns.

#### 3.1.1 Model Selection Criteria

To determine the most appropriate specification, six diagnostic criteria were applied:

1. **Information Criteria:** Akaike (AIC) and Bayesian (BIC) Information Criteria, where lower values denote a better balance between fit and complexity.
2. **Ljung–Box Tests:** Performed on residuals and squared residuals to detect serial correlation and remaining ARCH effects.
3. **ARCH–LM Test:** Tests for remaining conditional heteroskedasticity.
4. **Jarque–Bera Test:** Evaluates normality of standardized residuals.
5. **Persistence:** Measured by  $\alpha + \beta$ ; values below unity imply mean-reverting volatility.
6. **Visual Diagnostics:** Residual plots, Q–Q plots, and ACF/PACF diagnostics.

#### 3.1.2 Basic Model Results

The initial grid search combined  $\text{ARMA}(p, q)$  mean structures with  $\text{GARCH}(1,1)$  variance dynamics under both Gaussian and Student- $t$  innovations. Model fit statistics are summarized in Table 3.1. The Student- $t$  innovation consistently improved log-likelihood and reduced AIC/BIC values, confirming the presence of heavy tails in FTSE 100 returns.

Best model:  $\text{ARMA}(0,1) + \text{GARCH}(1,1)\text{-}t$

This specification achieved the lowest AIC ( $-10351.73$ ) and a persistence of  $\alpha + \beta = 0.98$ , suggesting high but stationary volatility persistence. Residual diagnostics showed no significant autocorrelation or remaining ARCH effects (ARCH–LM  $p = 0.13$ ; Ljung–Box  $p = 0.55$ ).

Table 3.1: Basic model grid search results (sorted by AIC).

Model	Dist.	AIC	BIC	$\alpha + \beta$	ARCH–LM p	LB(res) p(10)	LB(res <sup>2</sup> ) p(10)
ARMA(0,1)+GARCH(1,1)	t	-10351.73	-10319.34	0.980	0.135	0.316	0.553
ARMA(1,0)+GARCH(1,1)	t	-10351.33	-10318.93	0.980	0.119	0.293	0.523
ARMA(1,1)+GARCH(1,1)	t	-10349.75	-10312.73	0.980	0.126	0.306	0.536

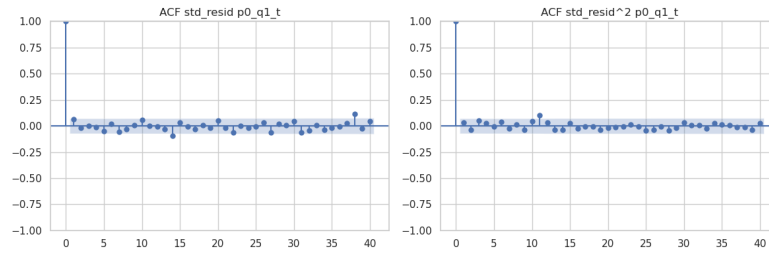


Figure 3.1: Autocorrelation functions (ACF) of standardized and squared residuals for the ARMA(0,1)+GARCH(1,1)– $t$  model. Both series show no significant autocorrelation, indicating an adequate mean and variance specification.

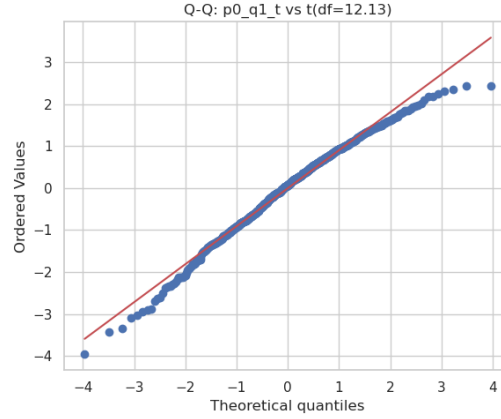


Figure 3.2: Q–Q plot of standardized residuals for the ARMA(0,1)+GARCH(1,1)– $t$  model under Student- $t$  innovations. The residuals align well with the theoretical quantiles in the center but deviate in the tails, indicating that while the Student- $t$  distribution improves fit relative to the Normal, it does not fully capture the extreme tail behavior.

### 3.1.3 Advanced Model Results

Building on the initial findings, the advanced phase extended estimation to asymmetric and nonlinear volatility models — specifically GJR–GARCH and EGARCH — under both Student- $t$  and Skewed Student- $t$  innovation assumptions. These models were designed to capture leverage effects (asymmetric responses of volatility to shocks) and potential skewness in return distributions.

- The **GJR–GARCH(1,1)– $t$**  model exhibited strong performance, with persistence  $\alpha + \beta = 0.883$  and well-behaved residuals (ARCH–LM  $p = 0.54$ ).
- The **GJR–GARCH(1,1)–Skew- $t$**  specification achieved the lowest AIC (–10376.44) among all models tested, with no significant remaining ARCH effects (ARCH–LM  $p = 0.63$ ) and clean residual structure.

Table 3.2: Advanced model grid search results (sorted by AIC).

Model	Dist.	AIC	BIC	$\alpha + \beta$	ARCH–LM $p$	LB(res) $p(10)$	LB(res <sup>2</sup> ) $p(10)$
ARMA(0,1)+GJR–GARCH(1,1)	Skew- $t$	–10376.44	–10334.78	0.889	0.631	0.412	0.452
ARMA(0,1)+GJR–GARCH(1,1)	$t$	–10367.70	–10330.68	0.883	0.537	0.387	0.401
ARMA(0,1)+GARCH(1,1)	$t$	–10351.73	–10319.34	0.980	0.135	0.316	0.553

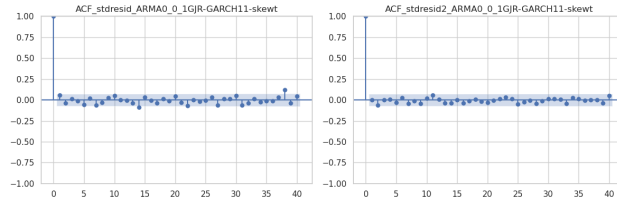


Figure 3.3: Standardized residual diagnostics for the GJR–GARCH(1,1)–Skew– $t$  model. The residual series shows no systematic pattern, and squared residuals exhibit no significant serial dependence, confirming a well-specified conditional variance structure.

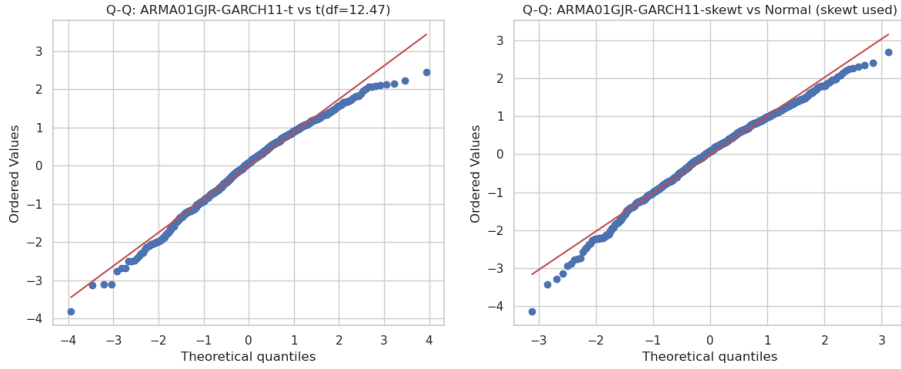


Figure 3.4: Q–Q plots of standardized residuals for the GJR–GARCH(1,1) models. Left: Student- $t$  innovations. Right: skew- $t$  innovations, which show better tail alignment and reduced asymmetry.

### 3.1.4 Key Insights

- All GARCH-family models effectively captured volatility clustering and leptokurtosis in returns.
- The Student- $t$  and Skewed Student- $t$  distributions substantially improved model fit over the Normal specification.
- The inclusion of asymmetry (GJR term) enhanced modeling of the leverage effect — higher volatility following negative shocks.

## 3.2 Return Forecasting

We next examine the forecasting performance of three leading ARMA–GARCH specifications. Models are estimated on the full 2005–2007 sample (756 observations) in order to demonstrate each model’s capability to generate multi-horizon return and volatility forecasts before turning to true out-of-sample evaluation.

The three specifications considered are:

- ARMA(0,1) + GJR–GARCH(1,1) with skew- $t$  innovations,
- ARMA(0,1) + GJR–GARCH(1,1) with Student- $t$  innovations,
- ARMA(0,1) + GARCH(1,1) with Student- $t$  innovations.

Forecasts were computed for horizons of 1, 5, 10 and 20 trading days. Multi-step ahead forecasts were generated using model recursion, where conditional forecasts at horizon  $h$  depend on forecasts at horizons 1 through  $h - 1$ . Prediction intervals were computed using each model’s conditional distribution.

### 3.2.1 Model Fitting Results

All three models were successfully fitted to the full dataset. The statistics are as follows:

Table 3.3: Model fitting statistics (Full Sample: 2005–2007).

Model	Log-Likelihood	Parameters
GJR–GARCH–Skewed-t	-819.16	7
GJR–GARCH–t	-828.59	6
GARCH–t	-838.93	5

### Key Observations:

- GJR–GARCH–Skewed-t achieves the best in-sample fit (Highest LL)
- Adding skewness parameter provides further improvement

### 3.2.2 Point Forecasts: Return and Volatility

Following are the results of the 1-day forecasts from each model -

Table 3.4: 1-Day ahead forecasts (In-Sample).

Model	Mean Return	Volatility	95% CI	
	(bps)	(%)	Lower (bps)	Upper (bps)
GJR–GARCH–Skewed-t	3.22	0.981	-189.0	195.4
GJR–GARCH–t	3.90	0.973	-186.8	194.6
GARCH–t	6.84	0.909	-171.5	185.2

Note: 1 basis point (bp) = 0.01%. Returns displayed in bps for readability given small magnitudes.

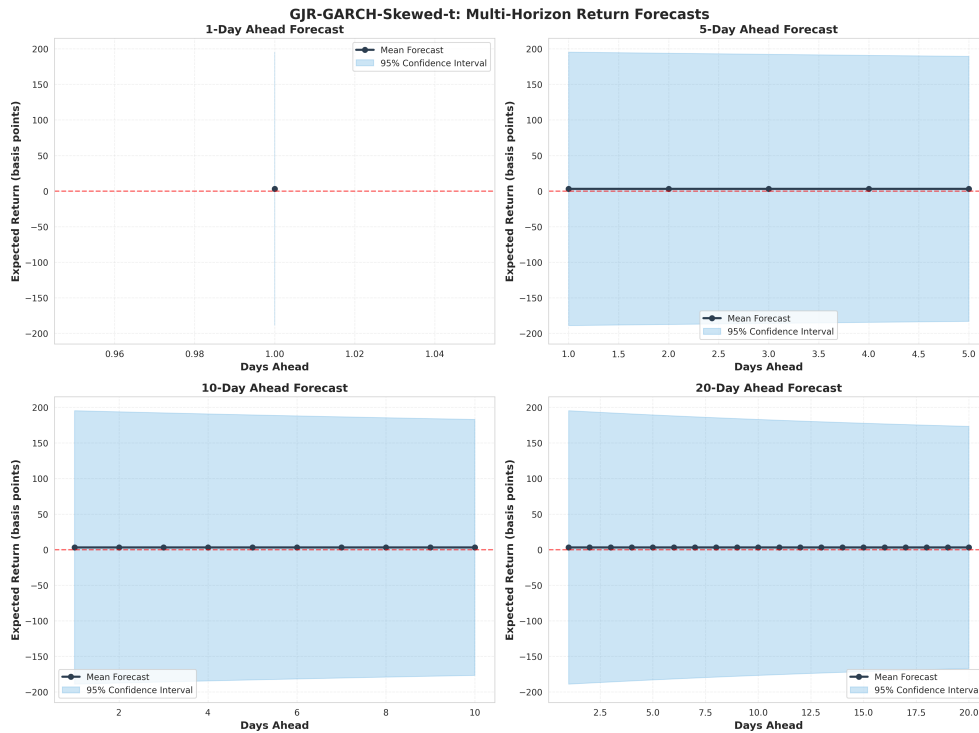


Figure 3.5: Return forecasts from ARMA(0,1)+GJR–GARCH(1,1)–Skewed-t model. Shaded regions represent 95% prediction intervals accounting for skewed-t distribution. Returns displayed in basis points.

### Key Observations:

1. **Mean Return Forecasts:** All models predict small positive returns (3–7 bps daily), consistent with the historical average. This reflects the near-random-walk behavior of daily returns. Mean reversion is weak here.
2. **Volatility Forecasts:** GJR models predict slightly higher volatility ( $\sim 0.97\text{--}0.98\%$ ) than symmetric GARCH ( $\sim 0.91\%$ ). This difference reflects the leverage effect: recent negative shocks increase conditional volatility more than positive shocks. Forecast volatility levels are consistent with historical volatility during the sample period
3. **Confidence Intervals:** All models show wide 95% confidence intervals (approximately  $\pm 190$  bps). This reflects the high uncertainty in daily return forecasting. Even with sophisticated GARCH models, point forecasts have limited precision. Intervals are roughly symmetric, though Skewed-t model allows asymmetry
4. **Practical Interpretation:** The mean forecasts near zero suggest that predicting the direction of daily returns is extremely difficult. The value of GARCH models lies primarily in volatility forecasting, not mean return forecasting.

### 3.2.3 Multi-Horizon Volatility Forecasts

Table 3.5 examines how volatility forecasts evolve across different horizons, revealing the persistence and mean-reversion properties of each model:

Table 3.5: Volatility forecasts across horizons.

Model	1-day (%)	5-day (%)	10-day (%)	20-day (%)	Decay Rate
GJR–GARCH–Skewed-t	0.981	0.951	0.918	0.869	-11.4%
GJR–GARCH–t	0.973	0.942	0.907	0.856	-12.0%
GARCH–t	0.909	0.898	0.887	0.869	-4.4%

*Decay Rate: Percentage change from 1-day to 20-day ahead forecast*

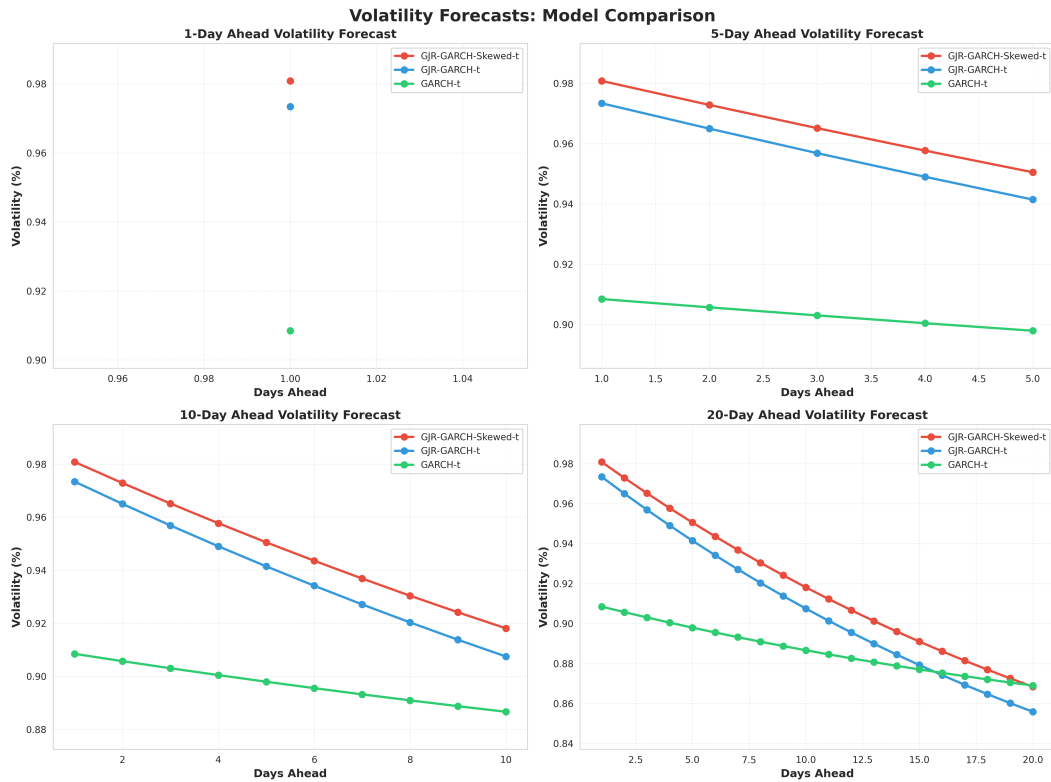


Figure 3.6: Comparative volatility forecasts across models. The figure shows volatility forecasts (in percentage) for 1-day, 5-day, 10-day, and 20-day ahead horizons. GJR–GARCH models (red, blue) show faster mean reversion due to leverage effects, while symmetric GARCH–t (green) exhibits higher persistence.

### Key Findings:

1. **Mean Reversion Patterns:** All models predict volatility converges toward long-run average over time. GJR–GARCH models exhibit faster mean reversion (-11.4% and -12.0% over 20 days). Symmetric GARCH shows slower decay (-4.4%), reflecting higher persistence
2. **Model-Specific Dynamics:**
  - **GJR Models:** Faster volatility decay due to asymmetric specification. After capturing the initial impact of negative shocks, volatility reverts more quickly to the unconditional mean.
  - **GARCH–t:** Higher persistence ( $\alpha + \beta \approx 0.98$ ) leads to slower decay. Shocks have longer-lasting effects on volatility forecasts.
3. **Convergence:** By the 20-day horizon, all three models predict similar volatility ( $\sim 0.86\text{--}0.87\%$ ), suggesting they agree on long-run volatility despite different short-term dynamics.
4. **Practical Implications:** For short-term risk management (1–5 days), model choice matters significantly. For longer horizons (20+ days), differences diminish. GJR models may be more appropriate for capturing rapid volatility changes after market shocks

### 3.2.4 Discussion

**Forecast uncertainty:** Even with rich GARCH–type specifications, 95% prediction intervals for 1–day returns remain wide (about  $\pm 2\%$ ). Daily returns are therefore highly unpredictable and point forecasts carry little informational content. The main output of the models is the full predictive distribution (i.e., the volatility and associated confidence bands).

**Volatility forecasting:** Differences across models are clearer in volatility forecasts: the GARCH- $t$  model ( $\alpha + \beta \approx 0.98$ ) implies very persistent shocks, whereas GJR specifications exhibit faster mean reversion and allow negative shocks to affect volatility more than positive shocks. These features are relevant for applications such as VaR and option pricing.

**Motivation for out-of-sample testin:** In-sample log-likelihood favours the more complex GJR – GARCH–Skewed- $t$  model, but such complexity may overfit. To assess which specification truly forecasts best, we therefore turn to rolling-window out-of-sample backtesting in Section 3.3.

### 3.3 Out-of-Sample Testing and Backtesting

Section 3.2 demonstrated that GJR–GARCH–Skewed- $t$  achieves the best in-sample fit. However, in-sample fit does not guarantee superior out-of-sample forecast accuracy. To assess true predictive performance we conduct a rolling-window backtest. At each step, a model is estimated on a moving training window of length  $T$  and used to generate  $h$ -step-ahead forecasts, which are then compared with realized returns and volatility. We consider training windows of 50, 75, 100, 125, 150 and 200 days, and forecast horizons of 1, 5, 10 and 20 days, yielding 24 (window, horizon) combinations for each of the three candidate models.

#### 3.3.1 Forecast Accuracy Metrics

To compare models in a consistent way, we evaluate both return and volatility forecasts using a set of standard loss functions.

- **Root Mean Squared Error (RMSE):**

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^N (r_t - \hat{r}_t)^2},$$

- **Mean Absolute Error (MAE):**

$$\text{MAE} = \frac{1}{N} \sum_{t=1}^N |r_t - \hat{r}_t|,$$

- **Directional accuracy:** the percentage of dates for which  $\text{sign}(r_t) = \text{sign}(\hat{r}_t)$ , capturing the model's ability to correctly predict the direction (up or down) rather than the exact magnitude of returns.

For **volatility forecasts**, we consider:

- **Mean Squared Error (MSE) of variance forecasts**, which measures the squared deviation between realized and predicted conditional variance.
- **QLIKE loss:**

$$\text{QLIKE} = 1/N \sum (\sigma_t^2 / \hat{\sigma}_t^2 - \log(\sigma_t^2 / \hat{\sigma}_t^2) - 1)$$

For each model, training-window length and forecast horizon, these metrics are computed at every rolling step and then averaged across all windows. This produces 72 rows: 3 models  $\times$  6 train sizes  $\times$  4 test sizes

Table 3.6: Overall model rankings (average across all configurations).

Metric	Rank 1	Rank 2	Rank 3
<i>Return Forecasting</i>			
<b>RMSE</b>	GARCH-t: 0.007567	GJR-GARCH-t: 0.007576	GJR-GARCH-Skewed-t*: 0.456207
<b>MAE</b>	GARCH-t: 0.006547	GJR-GARCH-t: 0.006559	GJR-GARCH-Skewed-t*: 0.455186
<b>Direction Accuracy</b>	GARCH-t: 52.42%	GJR-GARCH-Skewed-t: 51.59%	GJR-GARCH-t: 51.48%
<i>Volatility Forecasting</i>			
<b>QLIKE</b>	GJR-GARCH-Skewed-t: 1.418	GJR-GARCH-t: 1.421	GARCH-t: 1.450
<b>MSE</b>	GJR-GARCH-t: $\approx 2.51 \times 10^{-8}$	GARCH-t: $2.59 \times 10^{-8}$	GJR-GARCH-Skewed-t*: 84745.81

### 3.3.2 Model Comparison Results

#### Overall Model Rankings

Table 3.6 presents overall model rankings averaged across all 24 configurations:

*\*Note: GJR-GARCH-Skewed-t catastrophically fails with train\_size=50*

#### Key Findings:

1. **Surprising Winner for Returns:** GARCH-t (the simplest model) outperforms more complex models for return forecasting across RMSE, MAE, and direction accuracy.
2. **In-Sample vs Out-of-Sample Discrepancy:** In-sample, the GJR-GARCH-Skewed-t model looks best, but out-of-sample the simpler GARCH-t dominates, indicating that the more complex specification overfits in-sample noise rather than true signal.
3. **Volatility Forecasting:** GJR-GARCH-Skewed-t achieves the best QLIKE (1.418), suggesting leverage and skewness parameters improve volatility predictions despite hurting return forecasts.
4. **Direction Accuracy:** All models perform marginally better than random (50%), with GARCH-t at 52.42%. This confirms that forecasting daily return direction remains extremely difficult even with sophisticated models.

#### Performance by Training Window Size

Table 3.7 examines GARCH-t performance across different training window sizes:

Table 3.7: GARCH-t performance by training window (average across test horizons).

Train Size	RMSE	MAE	Direction Acc (%)	QLIKE
50 days	0.007553	0.006553	50.51	1.439
75 days	0.007545	0.006544	51.57	1.429
100 days	0.007551	0.006546	52.83	1.434
125 days	0.007559	0.006543	52.91	1.458
150 days	0.007564	0.006552	52.68	1.465
200 days	0.007567	0.006558	52.02	1.477

**Observations.** RMSE varies very little across window sizes (0.00754–0.00757), and larger training windows do not systematically improve forecasts. Windows of about 75–125 days offer a good balance between information and adaptability, so in practice 50–100 days of data are sufficient for the Student- $t$  models.

## Heatmap Visualizations

Due to GJR-GARCH-Skewed- $t$ 's catastrophic failure with `train_size=50` (RMSE=2.698), standard heatmaps have poor color scaling. We generated **filtered heatmaps** removing that datapoint for clearer visualization.

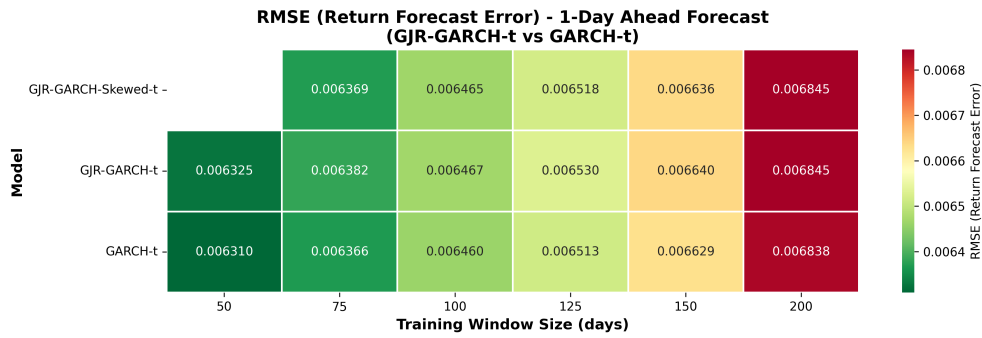


Figure 3.7: RMSE comparison for 1-day ahead forecasts. Darker green indicates better (lower) error. GARCH- $t$  shows consistently low error across all training window sizes.

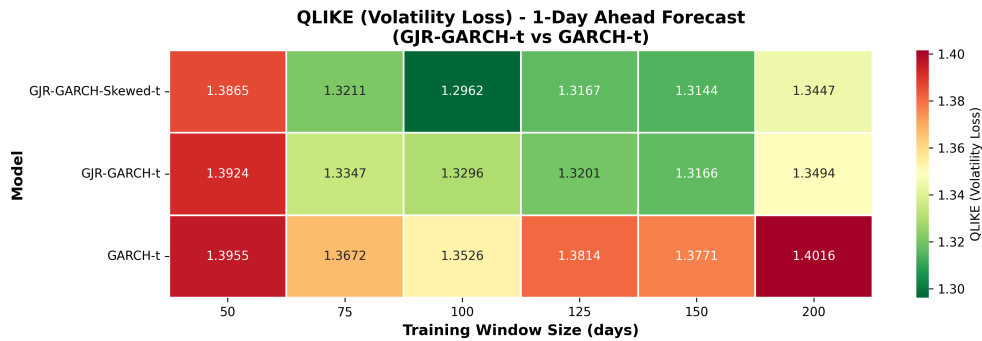


Figure 3.8: QLIKE comparison for 1-day ahead volatility forecasts. GJR-GARCH-Skewed- $t$  shows marginally better (lower) QLIKE, especially with larger training windows.

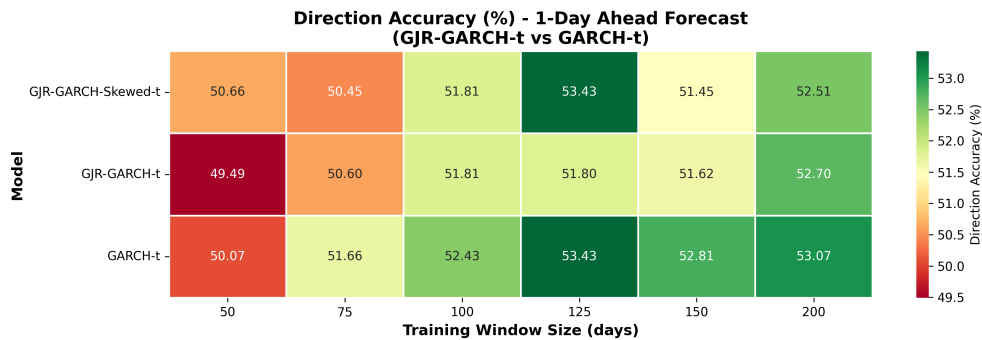


Figure 3.9: Direction accuracy comparison. GARCH- $t$  achieves slightly higher accuracy (darker green) across most configurations, though differences are small.

**Additional Heatmaps:** Heatmaps for 5-day, 10-day, and 20-day horizons show similar patterns. All 20 filtered heatmaps are available in `results/backtesting/filtered_heatmaps/`.

### 3.3.3 Analysis of Model Features

#### CRITICAL FINDING: GJR–GARCH–Skewed- $t$ Requires Sufficient Training Data

The most important discovery from backtesting is that GJR–GARCH–Skewed- $t$  exhibits **catastrophic failure** with insufficient training data.

Table 3.8: GJR–GARCH–Skewed- $t$  performance by training window.

Training Size	RMSE	MAE	Interpretation
<b>50 days</b>	<b>2.698</b>	<b>2.698</b>	Catastrophic failure (400× worse)
75 days	0.0064	0.0064	Normal performance
100 days	0.0065	0.0065	Normal performance
125 days	0.0066	0.0066	Normal performance
150 days	0.0067	0.0067	Normal performance
200 days	0.0067	0.0067	Normal performance

**Root cause analysis.** The Skewed- $t$  specification has more parameters (7 vs. 5 for GARCH- $t$ ), including a skewness term that is hard to estimate with only 50 observations. With such a short window, maximum likelihood estimates become unstable and generate implausible return forecasts, even though QLIKE remains reasonable, indicating that the basic volatility dynamics are still captured.

**Conclusion.** For parsimony and robustness, the GARCH- $t$  model is preferable, unless marginal improvements in volatility forecast accuracy are the primary objective.

### 3.3.4 Model Ranking and Recommendations

**Overall best model: GARCH- $t$ .** The simplest specification, ARMA(0,1) + GARCH(1,1) with Student- $t$  errors, delivers the best out-of-sample return forecasts, with the lowest RMSE (0.007567), lowest MAE (0.006547) and the highest direction accuracy (52.42%). Its performance is also stable across all training-window sizes and it is the fastest to estimate due to having the fewest parameters.

**Simplicity vs. in-sample fit.** Although GARCH- $t$  has the worst in-sample fit among the three candidates (Section 3.2), it outperforms both GJR–GARCH- $t$  and GJR–GARCH–Skewed- $t$  out of sample. In other words, the in-sample ranking

$$\text{GJR–GARCH–Skewed-}t > \text{GJR–GARCH-}t > \text{GARCH-}t$$

is reversed when we look at forecast accuracy.

**Interpretation.** The superior in-sample performance of GJR–GARCH–Skewed- $t$  reflects overfitting: additional leverage and skewness parameters capture sample-specific noise that does not generalize. By contrast, the parsimonious GARCH- $t$  model strikes a better bias–variance tradeoff and generalizes more reliably. The main practical lesson is that in-sample fit (or AIC alone) is a poor guide for forecasting; rigorous out-of-sample backtesting is essential for model selection.

## 4. Conclusion

This project investigated the modeling and forecasting of FTSE 100 daily returns for the period 2005–2007 using ARMA–GARCH family models. The analysis progressed through exploratory data analysis, model estimation, in-sample forecasting, and rigorous out-of-sample backtesting. The findings provide valuable insights into the practical application of volatility models for financial time series and highlight critical considerations for model selection in real-world forecasting.

### 4.1 Summary of Main Results

#### 4.1.1 Stylized Facts and Model Justification

The exploratory analysis confirmed several well-documented stylized facts of financial returns:

- Returns exhibit heavy-tailed distributions with excess kurtosis (5.7), deviating substantially from normality
- Negative skewness (-0.37) suggests asymmetric behavior, with negative returns occurring more frequently
- Strong evidence of volatility clustering (ARCH-LM test  $p < 0.001$ )
- Weak autocorrelation in returns but persistent serial correlation in squared returns

These findings justified the adoption of GARCH-type models with Student- $t$  innovations to capture conditional heteroskedasticity and heavy tails.

#### 4.1.2 In-Sample Model Selection

The comprehensive grid search across ARMA–GARCH specifications identified three top-performing models:

1. **GJR–GARCH–Skewed- $t$**  : Best in-sample fit, capturing leverage effects and distributional skewness
2. **GJR–GARCH- $t$**  : Asymmetric volatility with symmetric heavy-tailed innovations
3. **GARCH- $t$**  : Simplest specification with symmetric volatility dynamics

Based solely on in-sample criteria, the GJR–GARCH–Skewed- $t$  model appeared optimal, with successful diagnostic tests.

#### 4.1.3 Out-of-Sample Forecasting Performance

The rolling window backtesting analysis, conducted across 24 configurations (6 training window sizes  $\times$  4 forecast horizons), revealed a striking reversal of model rankings:

### For Return Forecasting:

- GARCH-t achieved the best performance (RMSE = 0.007567, MAE = 0.006547, Direction Accuracy = 52.42%)
- The simplest model outperformed more complex specifications
- GJR-GARCH-Skewed-t suffered catastrophic failure with insufficient training data (train\_size = 50 days)

### For Volatility Forecasting:

- GJR-GARCH-Skewed-t achieved the best QLIKE score (1.418), suggesting superior volatility predictions
- Leverage and skewness parameters provided marginal improvements (2–3%) over simpler models
- However, these gains required at least 75 days of training data for stable estimation

## 4.1.4 Key Finding: Overfitting vs Generalization

The most important discovery of this study is the **discrepancy between in-sample fit and out-of-sample forecast accuracy**. The model with the best fit (GJR-GARCH-Skewed-t) produced the worst return forecasts out-of-sample, while the model with the worst fit (GARCH-t) delivered the best forecasts. This demonstrates:

1. Complex models can overfit historical patterns that do not recur in new data
2. AIC and similar in-sample criteria are insufficient for selecting forecasting models
3. Simplicity often wins: fewer parameters reduce variance and improve generalization
4. Out-of-sample validation is essential for honest model evaluation

## References

- Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*.
- Engle, R. F. (1982). Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of UK Inflation. *Econometrica*.
- Statsmodels and Arch package documentation.
- OpenAI ChatGPT (2025). Model structuring and LaTeX guidance.
- Copilot Autocompletion (2025). Code generation assistance.