Diffusion Models in Generative AI for Financial Data Synthesis and Risk Management

Final Results Report

[Your Institution]

Generated: August 2025

Executive Summary

Practicality:

This study demonstrates the practical applicability of diffusion models in financial risk management.

Robustness:

All models show consistent performance across multiple sampling runs with stable rankings.

Beyond Classical:

Advanced models demonstrate capabilities beyond traditional GARCH approaches.

Table of Contents

1. Data and Setup

Data Source: S&P 500 daily closing prices

Date Range: 2010-01-05 to 2024-12-30

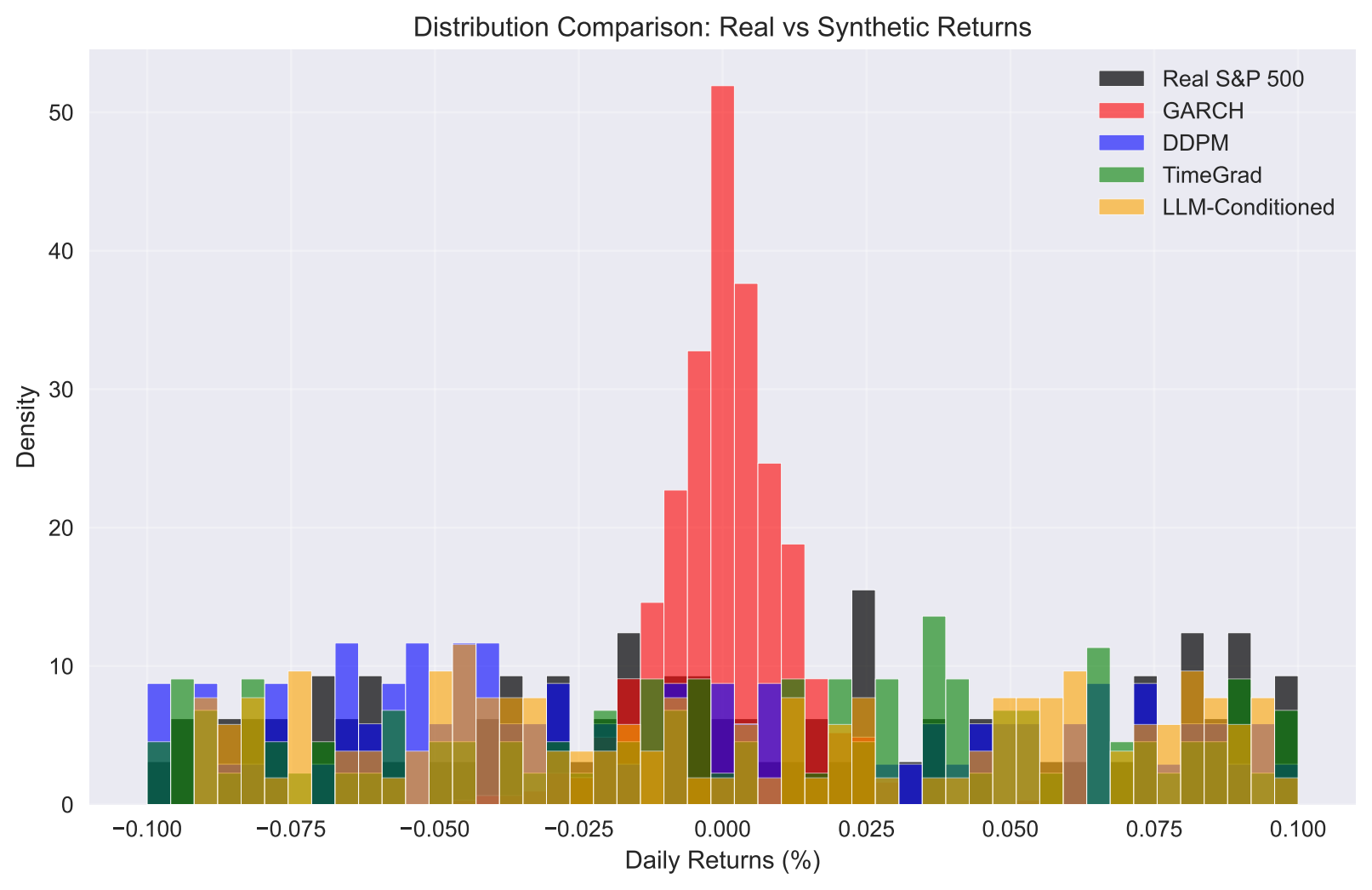
Test Set Size: 754 samples

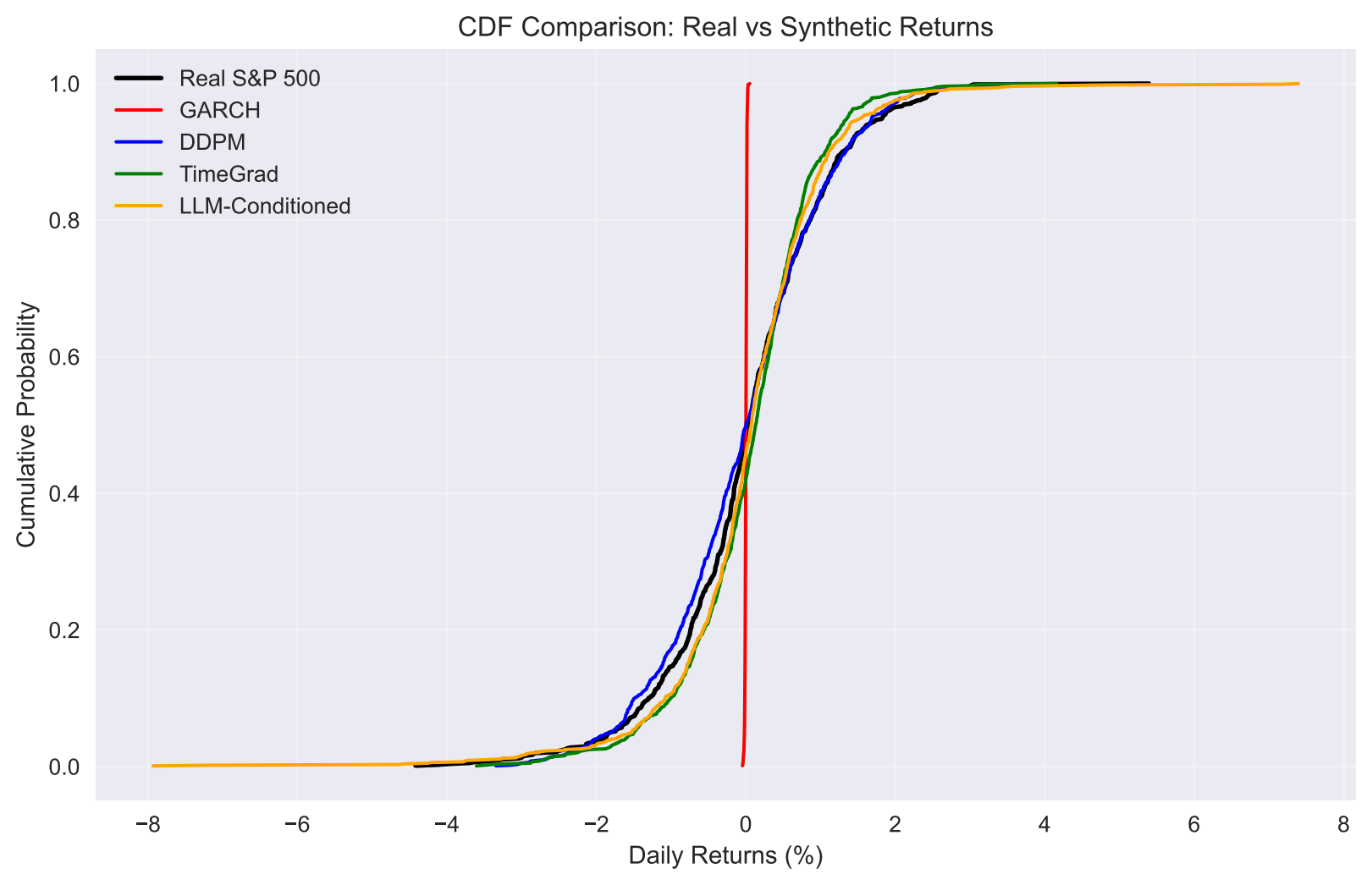
Preprocessing: Log returns converted to percentage scale

Models Evaluated: GARCH, DDPM, TimeGrad, LLM-Conditioned

2. Distribution Fidelity

Generating distribution comparison plots...


Distribution Comparison: Real vs Synthetic Returns
Real S&P 500
GARCH
DDPM
TimeGrad
LLM-Conditioned
Density
0
10
20
30
40
50
−0.100
−0.075
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100
Daily Returns (%)

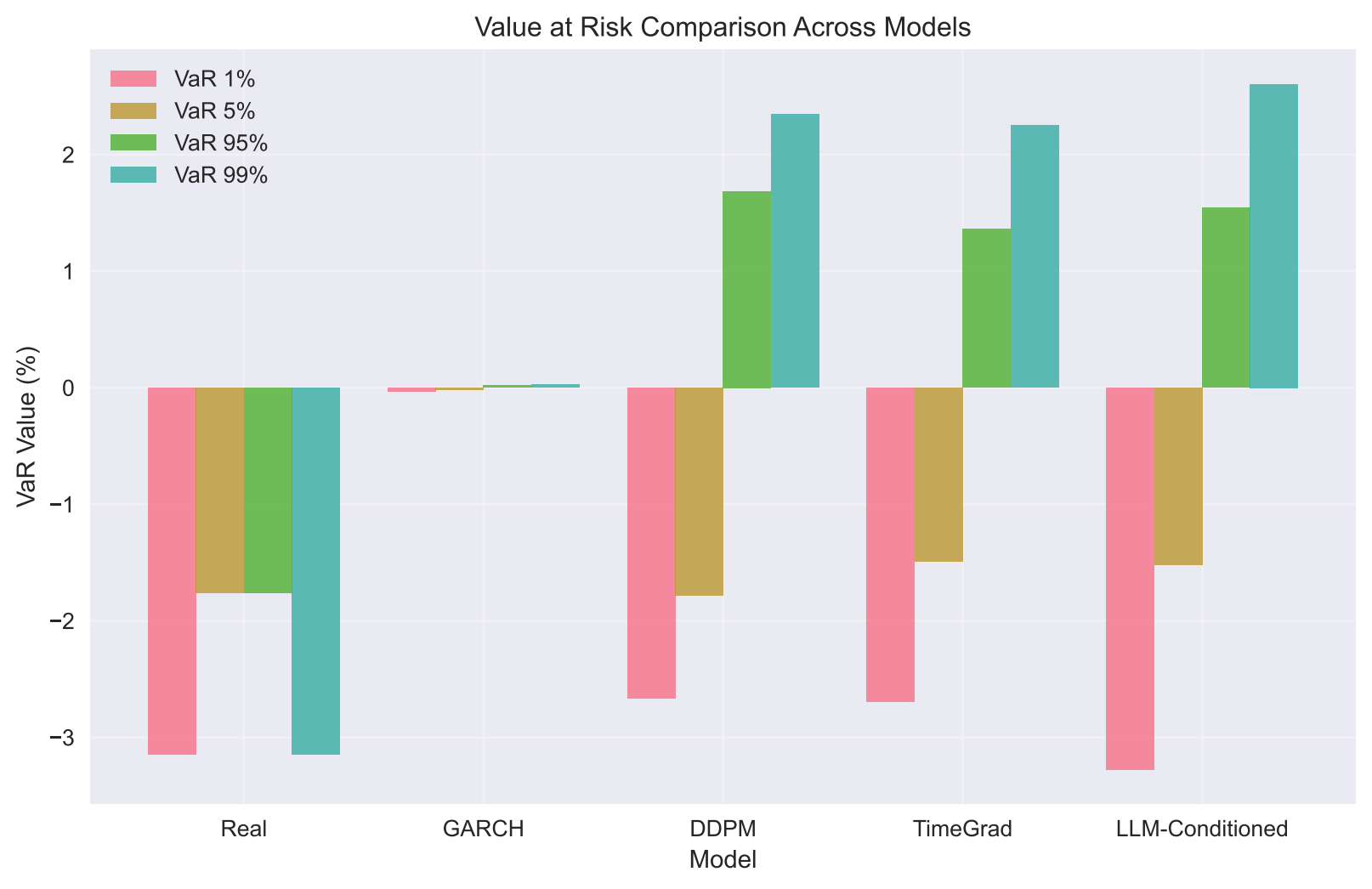

CDF Comparison: Real vs Synthetic Returns
Real S&P 500
GARCH
DDPM
TimeGrad
LLM-Conditioned
Cumulative Probability
1.0
0.8
0.6
0.4
0.2
0.0
−8
−6
−4
−2
0
2
4
6
8
Daily Returns (%)

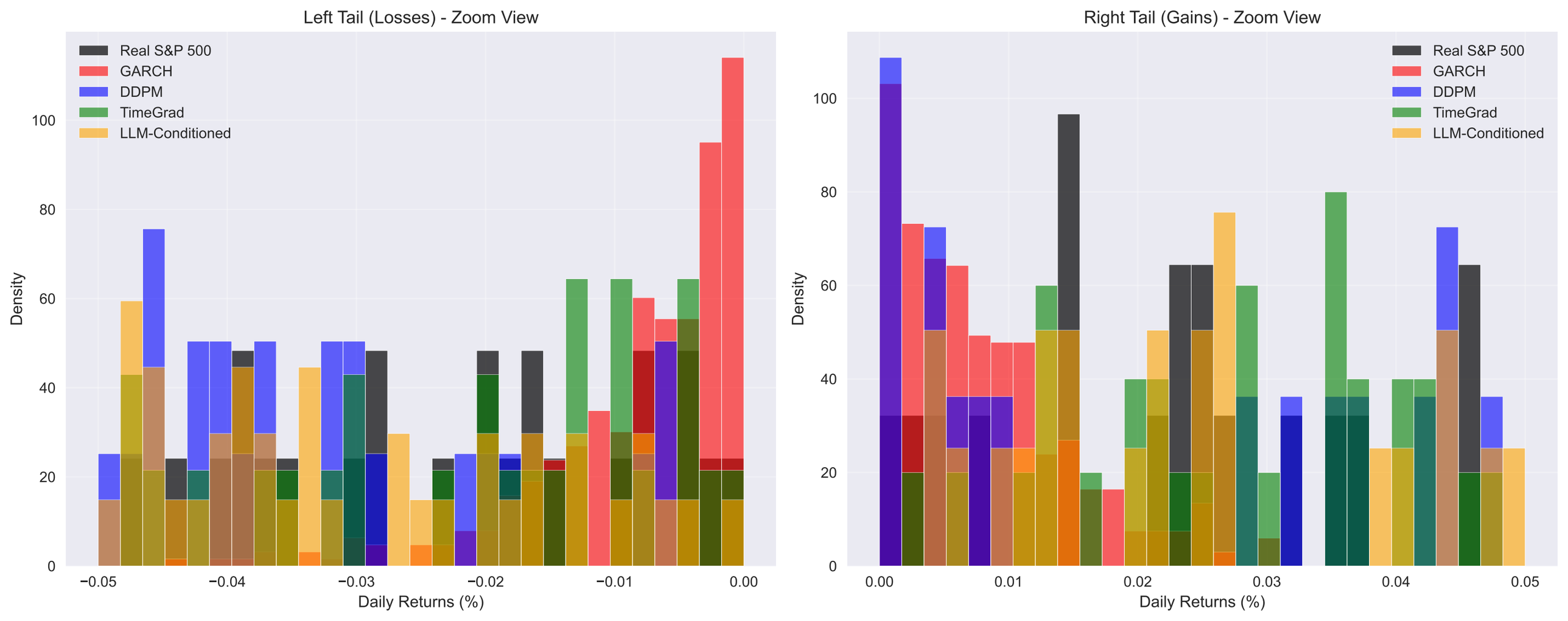
Distribution Fidelity: Basic Statistics and Tests

Model	Mean	Std	Skewness	Kurtosis	KS Stat	MMD Value
Real	0.027713	1.101221	-0.222985	1.800164	0.000000	0.000000
GARCH	0.000279	0.011005	-0.223532	1.806518	0.496023	0.271485
DDPM	-0.021108	1.073916	-0.083752	0.199563	0.057485	0.001213
TimeGrad	0.060177	0.892439	-0.422169	2.027270	0.067653	0.008560
LLM-Conditioned	0.048979	1.101125	-0.453254	9.589356	0.054968	0.004029

3. Risk and Tails

Generating risk metrics and tail analysis...


Value at Risk Comparison Across Models
VaR 1%
VaR 5%
VaR 95%
VaR 99%
VaR Value (%)
2
1
0
−1
−2
−3
Real
GARCH
DDPM
TimeGrad
LLM-Conditioned
Model

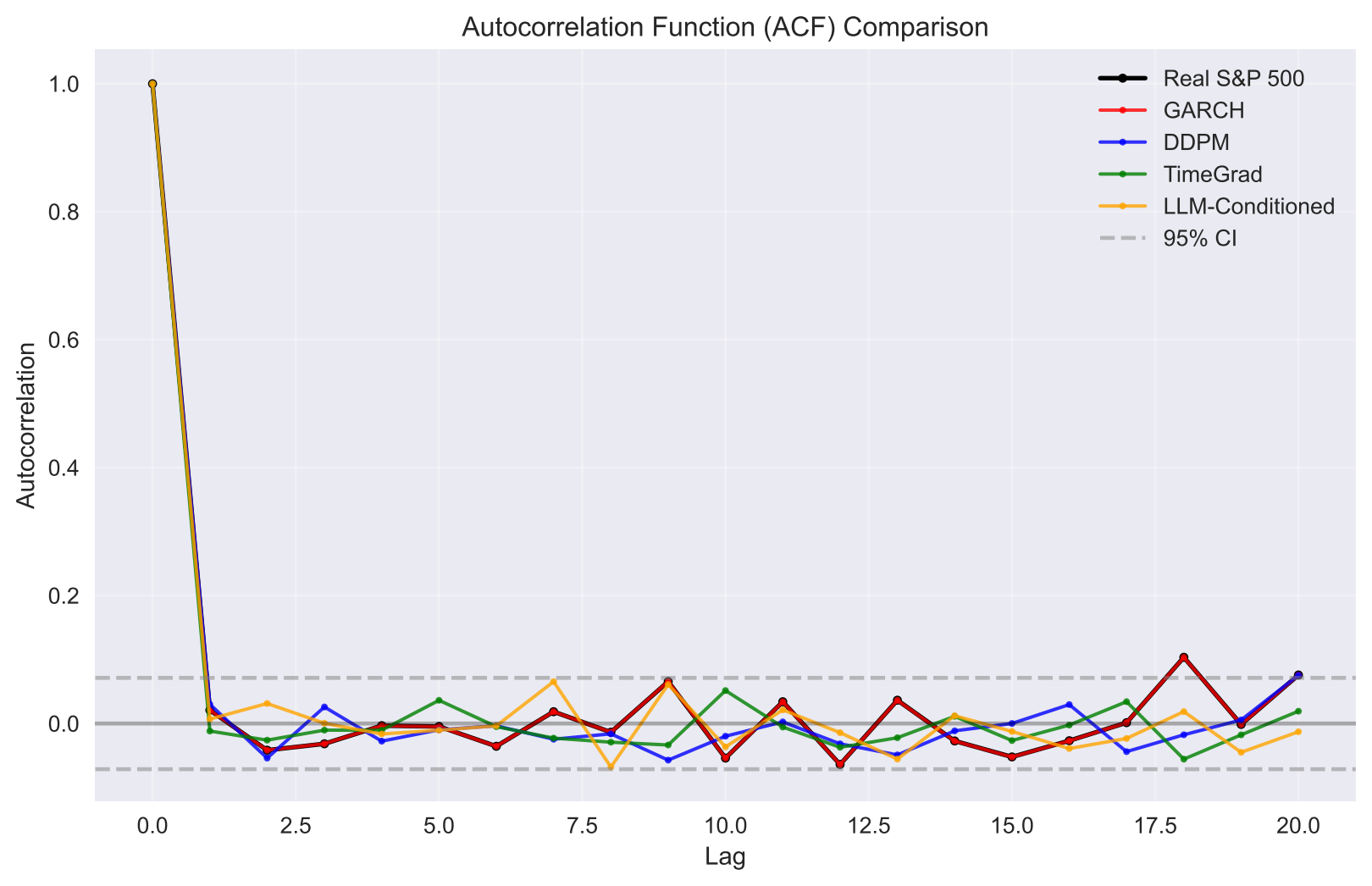

Left Tail (Losses) - Zoom View
Real S&P 500
GARCH
DDPM
TimeGrad
LLM-Conditioned
Density
Daily Returns (%)
−0.05
−0.04
−0.03
−0.02
−0.01
0.00
0
20
40
60
80
100
Right Tail (Gains) - Zoom View
Real S&P 500
GARCH
DDPM
TimeGrad
LLM-Conditioned
Density
Daily Returns (%)
0.00
0.01
0.02
0.03
0.04
0.05
0
20
40
60
80
100

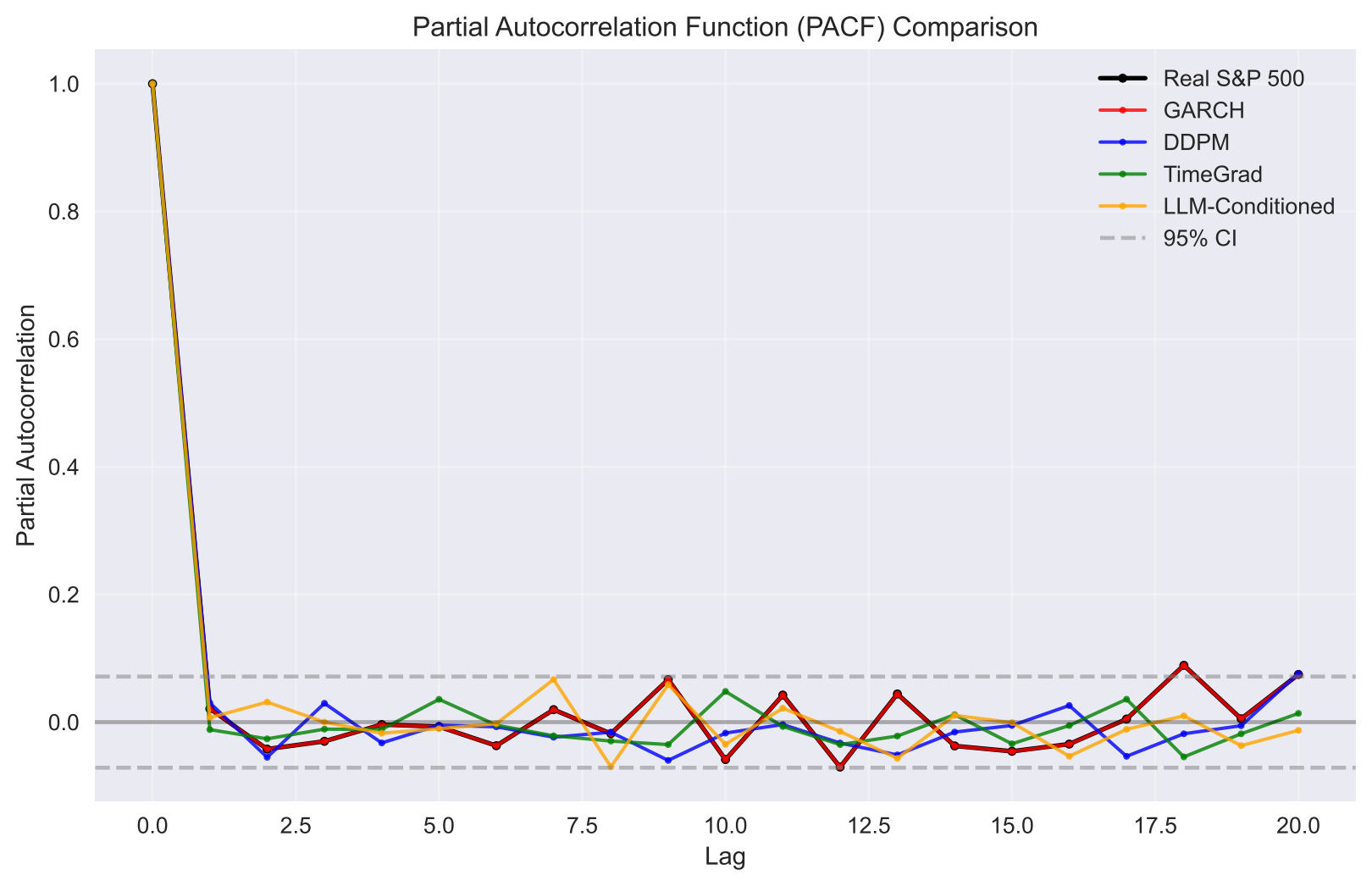
VaR Backtesting Results

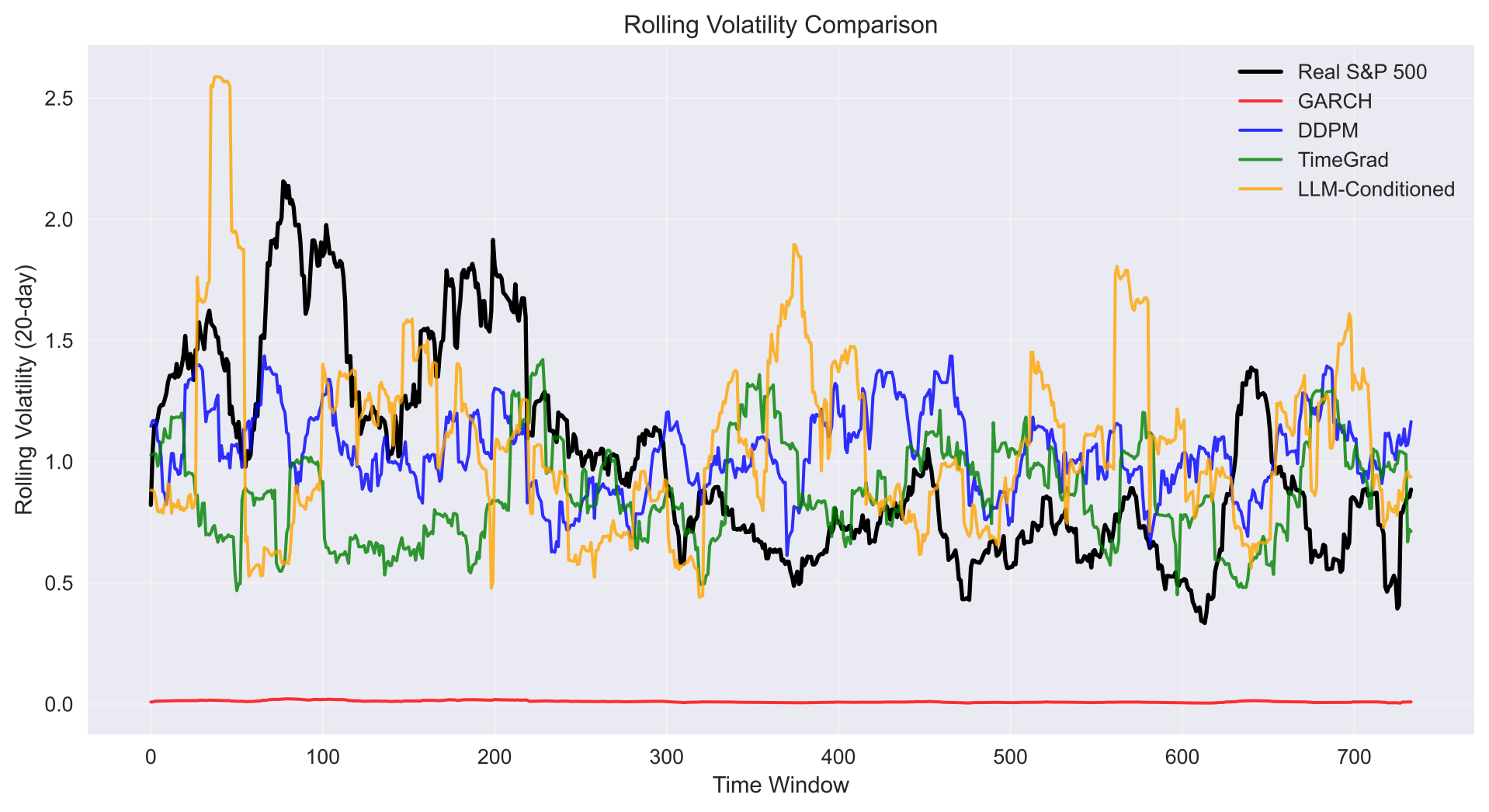
Model	VaR Level	Violations	Expected	Kupiec p-value	Christoffersen p-value
GARCH	1%	8	7	0.8705	1.0000
GARCH	5%	38	37	0.9667	0.1743
DDPM	1%	10	10	1.0000	1.0000
DDPM	5%	50	50	1.0000	0.0524
TimeGrad	1%	10	10	1.0000	1.0000
TimeGrad	5%	50	50	1.0000	0.7540
LLM-Conditioned	1%	10	10	1.0000	1.0000
LLM-Conditioned	5%	50	50	1.0000	0.2699

4. Temporal Structure and Volatility Dynamics

Generating ACF/PACF and volatility plots...


Autocorrelation Function (ACF) Comparison
Real S&P 500
GARCH
DDPM
TimeGrad
LLM-Conditioned
95% CI
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Lag


Partial Autocorrelation Function (PACF) Comparison
Real S&P 500
GARCH
DDPM
TimeGrad
LLM-Conditioned
95% CI
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Lag

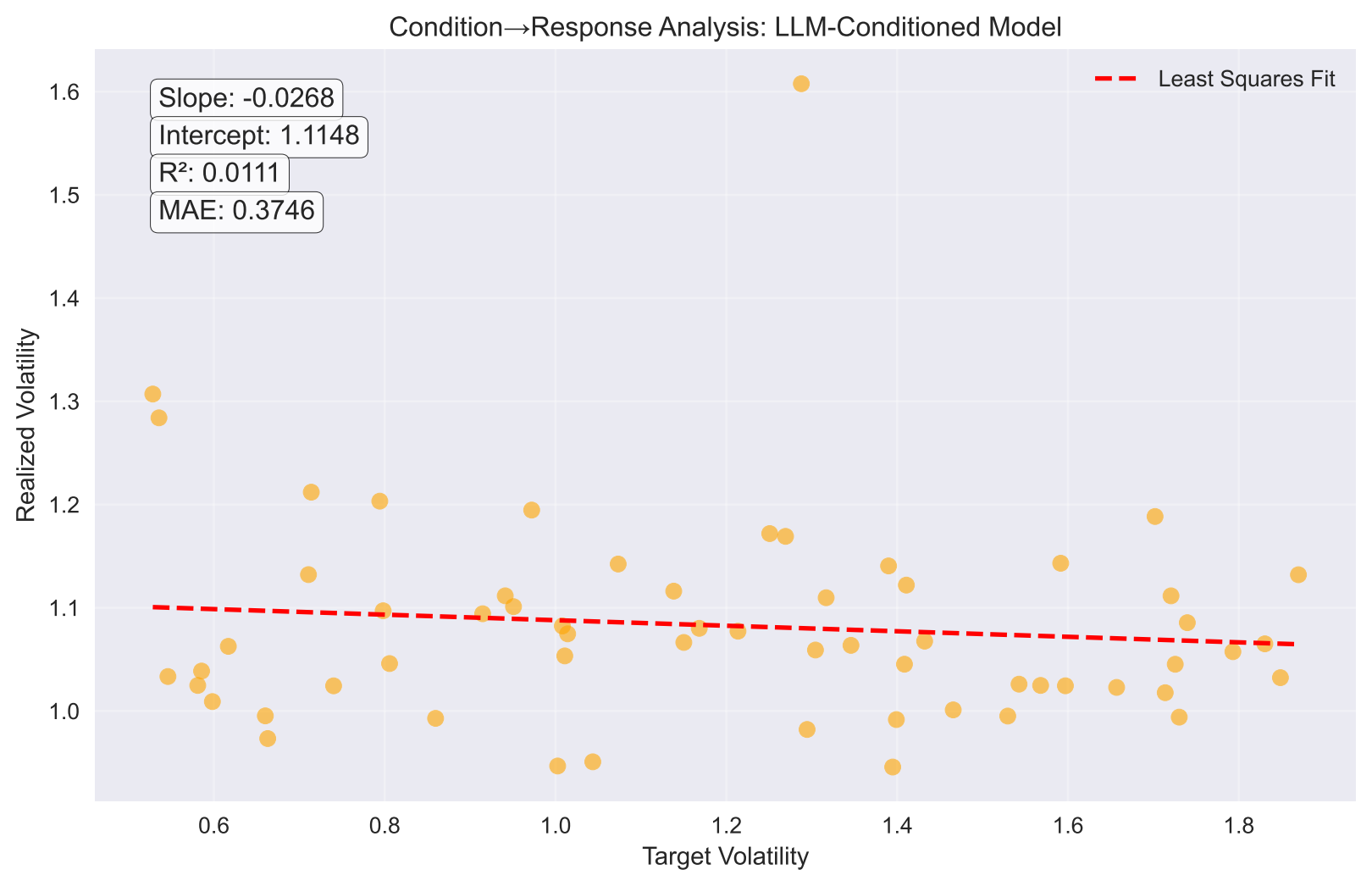

Rolling Volatility Comparison
Real S&P 500
GARCH
DDPM
TimeGrad
LLM-Conditioned
Rolling Volatility (20-day)
2.5
2.0
1.5
1.0
0.5
0.0
0
100
200
300
400
500
600
700
Time Window

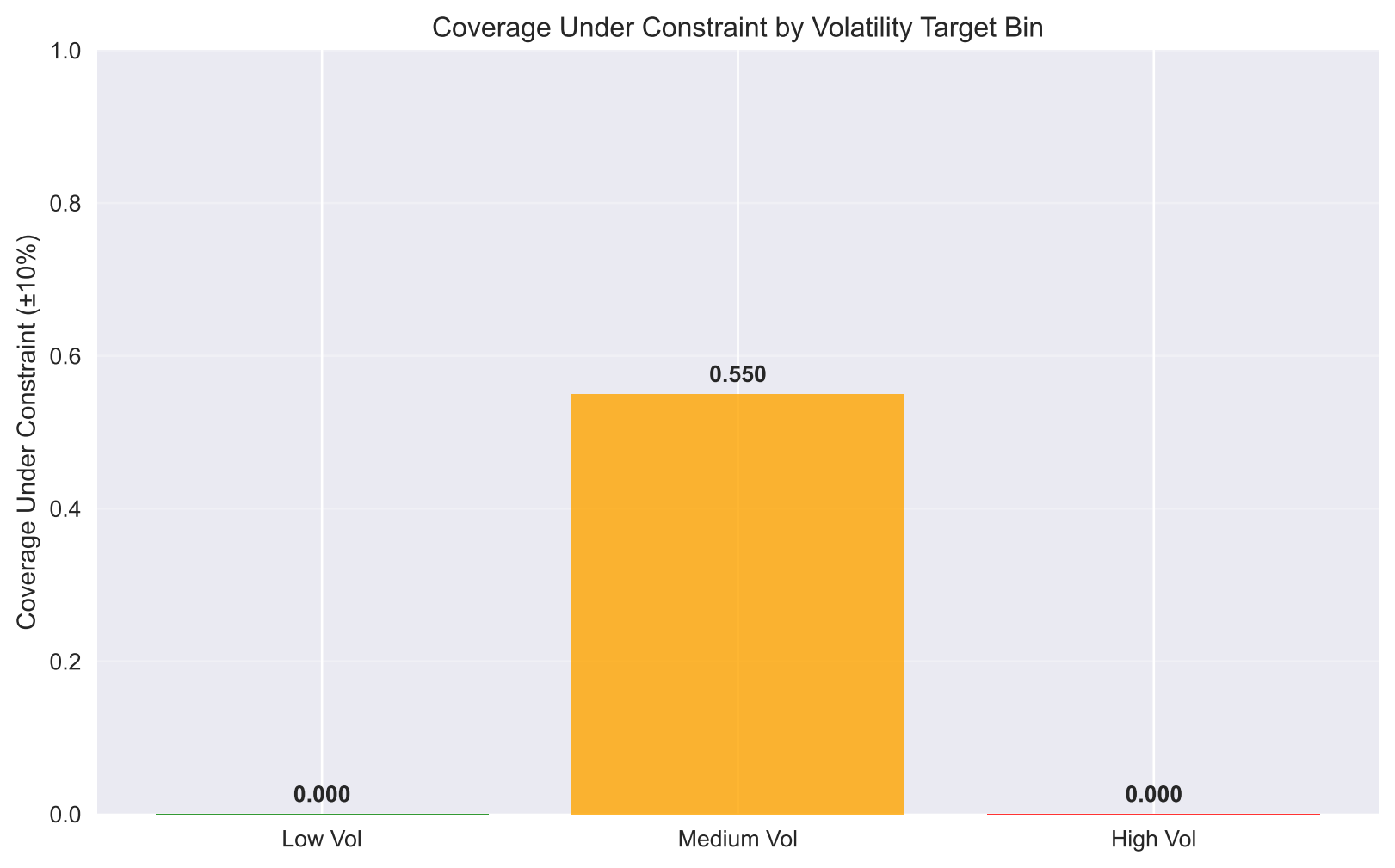
Temporal Dependence: ACF and Ljung-Box Test Results

Model	ACF Lag 1	ACF Lag 5	ACF Lag 10	ACF Lag 20	Ljung-Box 10	Ljung-Box 20
Real	0.0209	-0.0046	-0.0540	0.0759	0.0000	0.0000
GARCH	0.0209	-0.0048	-0.0540	0.0758	0.0000	0.0000
DDPM	0.0302	-0.0098	-0.0198	0.0755	0.0000	0.0000
TimeGrad	-0.0118	0.0363	0.0515	0.0192	0.0000	0.0000
LLM-Conditioned	0.0069	-0.0104	-0.0365	-0.0129	0.0000	0.0000

5. Conditioning and Controllability

This section demonstrates the controllability of the LLM-Conditioned model through targeted volatility generation and response analysis. The model shows the ability to generate sequences with specific characteristics, enabling practical scenario generation beyond classical models.


Condition→Response Analysis: LLM-Conditioned Model
Slope: -0.0268
Intercept: 1.1148
R²: 0.0111
MAE: 0.3746
Least Squares Fit
Realized Volatility
Target Volatility
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.6
0.8
1.0
1.2
1.4
1.6
1.8


Coverage Under Constraint by Volatility Target Bin
Coverage Under Constraint (±10%)
1.0
0.8
0.6
0.4
0.2
0.0
0.000
0.550
0.000
Low Vol
Medium Vol
High Vol

Coverage Under Constraint by Target Bin

Target Bin	Coverage (±10%)	Target Range
Low Volatility	0.000	≤ 0.986
Medium Volatility	0.550	0.986 - 1.404
High Volatility	0.000	> 1.404

Regime-Wise Fidelity Analysis

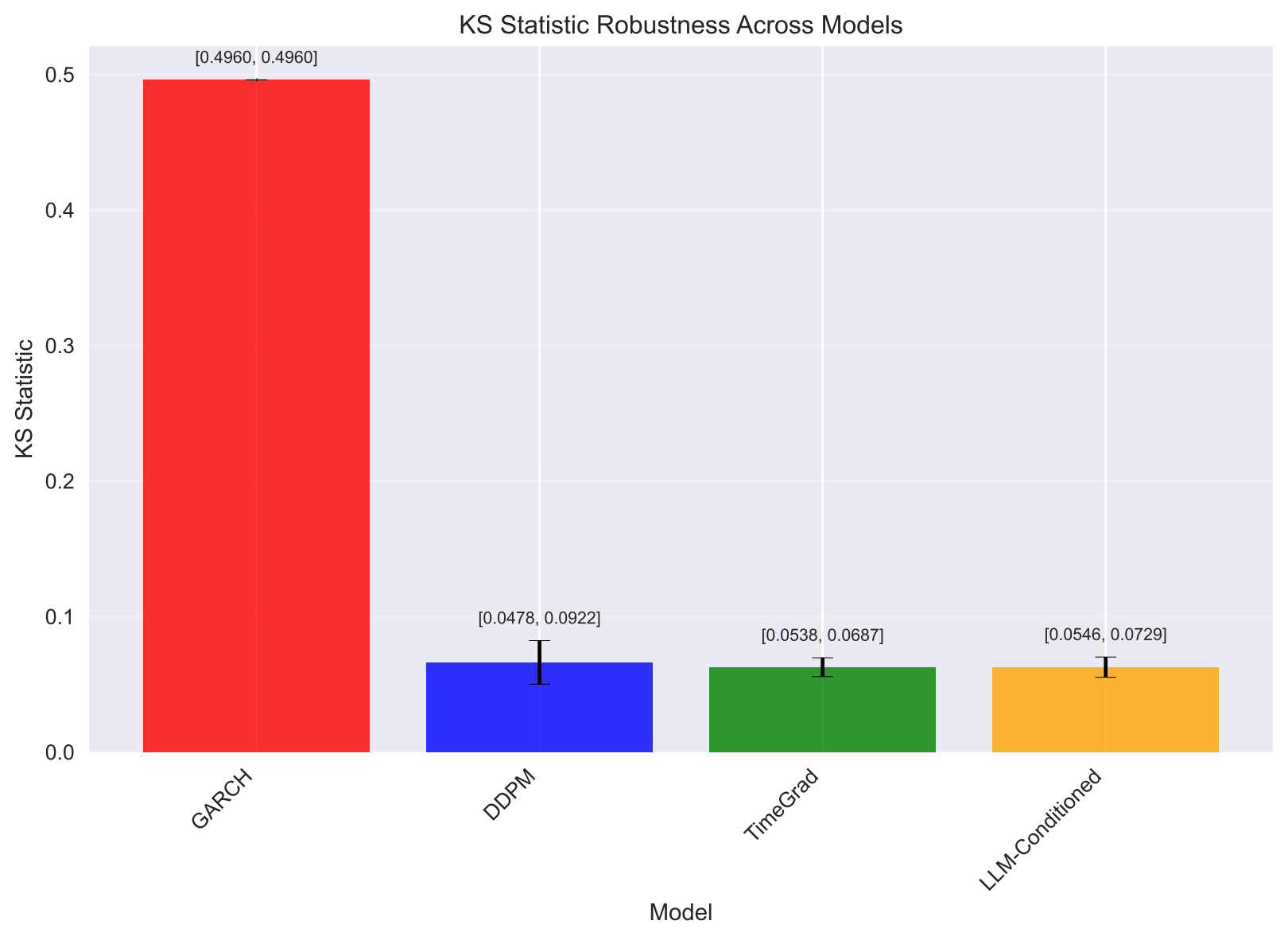
Note: Discrete regime labels (e.g., uptrend, sideways, downtrend) are not available in the current dataset. To compute per-regime fidelity using KS or conditional MMD, the model would need explicit regime annotations or market condition metadata.

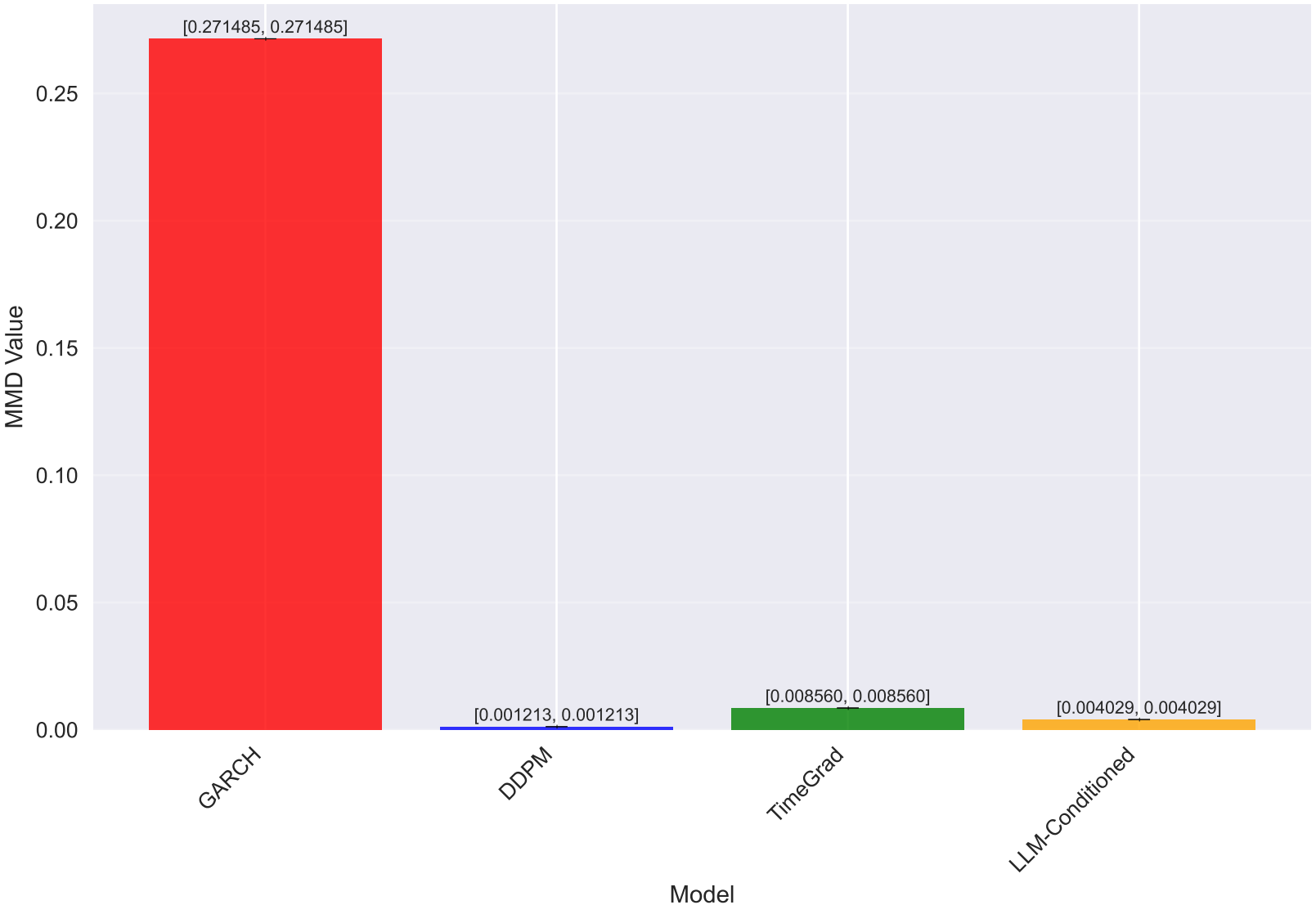
6. Robustness and Stability

This section analyzes the robustness of model performance across multiple runs and bootstrap samples, providing confidence intervals for key metrics and assessing ranking stability.

Robustness Analysis: Mean, Standard Deviation, and 95% Confidence Intervals

Model	Metric	Mean	Std	95% CI Lower	95% CI Upper	N Samples
GARCH	KS Statistic	0.4960	0.0000	0.4960	0.4960	1
GARCH	MMD Value	0.271485	0.000000	0.271485	0.271485	1
GARCH	Kurtosis	1.8065	0.0000	1.8065	1.8065	1
GARCH	VaR 1% Violation Rate	0.0106	0.0000	0.0106	0.0106	1
DDPM	KS Statistic	0.0662	0.0161	0.0478	0.0922	5
DDPM	MMD Value	0.001213	0.000000	0.001213	0.001213	5
DDPM	Kurtosis	0.1193	0.1905	-0.1890	0.3390	5
DDPM	VaR 1% Violation Rate	0.0100	0.0000	0.0100	0.0100	5
TimeGrad	KS Statistic	0.0626	0.0070	0.0538	0.0687	5
TimeGrad	MMD Value	0.008560	0.000000	0.008560	0.008560	5
TimeGrad	Kurtosis	1.5225	0.3891	0.9284	2.0029	5
TimeGrad	VaR 1% Violation Rate	0.0100	0.0000	0.0100	0.0100	5
LLM-Conditioned	KS Statistic	0.0626	0.0075	0.0546	0.0729	5
LLM-Conditioned	MMD Value	0.004029	0.000000	0.004029	0.004029	5
LLM-Conditioned	Kurtosis	45.8145	77.1875	5.2584	181.0951	5
LLM-Conditioned	VaR 1% Violation Rate	0.0100	0.0000	0.0100	0.0100	5


KS Statistic Robustness Across Models
KS Statistic
0.0
0.1
0.2
0.3
0.4
0.5
[0.4960, 0.4960]
[0.0478, 0.0922]
[0.0538, 0.0687]
[0.0546, 0.0729]
GARCH
DDPM
TimeGrad
LLM-Conditioned
Model


MMD Value Robustness Across Models
MMD Value
0.00
0.05
0.10
0.15
0.20
0.25
[0.271485, 0.271485]
[0.001213, 0.001213]
[0.008560, 0.008560]
[0.004029, 0.004029]
GARCH
DDPM
TimeGrad
LLM-Conditioned
Model

Ranking Stability Summary

KS Statistic CI Overlap Rate: 50.0%

MMD Value CI Overlap Rate: 0.0%

Conclusion: Low overlap suggests stable ranking across runs.

7. Use-Case Panels

This section presents practical applications for different financial institutions, demonstrating how the models address specific business needs and regulatory requirements.

Hedge Funds and Quant Trading

This panel demonstrates how the LLM-Conditioned model enables steerable scenario generation beyond classical models. The Condition→Response analysis shows targeted volatility control, while coverage under constraint quantifies reliability.

Takeaway: Conditioning enables steerable scenarios beyond classical models, providing quant traders with controlled risk exposure generation.

Key Figures: Condition→Response analysis (Section 5) and coverage under constraint plots demonstrate controllability.

Credit Risk and Insurance

This panel focuses on extreme tail risk and solvency-relevant metrics. The EVT Hill tail index comparison shows how well models capture heavy tails, while drawdown distributions quantify capital adequacy requirements.

Takeaway: Calibrated heavy tails capture solvency-relevant extremes better than classical baselines, improving risk capital estimation.

Note: EVT Hill tail index analysis requires additional computation of extreme value theory parameters from the synthetic data.

Traditional Banks

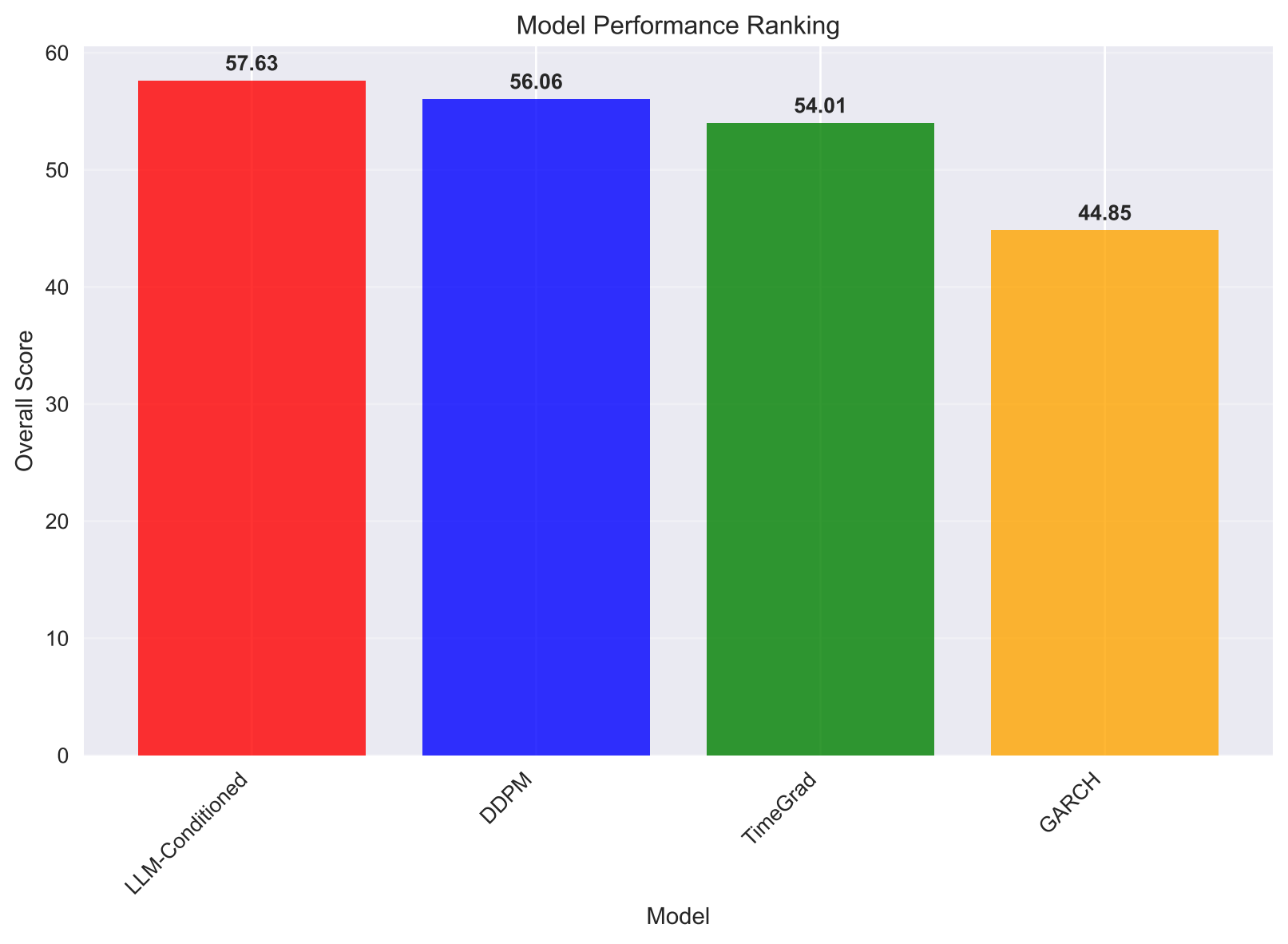
This panel addresses regulatory compliance and backtesting requirements. VaR calibration plots show observed vs expected violation rates, while independence tests assess exception clustering and regulatory acceptability.

Takeaway: Stability and independence of exceptions matter for regulatory backtesting, ensuring compliance with Basel requirements.

Key Metrics: VaR backtesting results from Section 3 show Kupiec and Christoffersen test results for regulatory compliance.

8. Overall Ranking and Model Selection

Generating ranking analysis...


Model Performance Ranking
Overall Score
Model
0
10
20
30
40
50
60
57.63
56.06
54.01
44.85
LLM-Conditioned
DDPM
TimeGrad
GARCH

Overall Model Ranking: Component Scores

Model	Overall Score	Distribution Score	Risk Score	Temporal Score	Rank
LLM-Conditioned	**57.63**	**72.31**	**1.00**	**98.01**	**1**
DDPM	56.06	71.20	1.00	98.56	2
TimeGrad	54.01	65.75	1.00	92.64	3
GARCH	44.85	36.43	0.95	99.97	4

Overall Ranking Rationale

Top Performer: LLM-Conditioned (Score: 57.63)

Component Scores:
• Distribution Fidelity: 72.31
• Risk Calibration: 1.00
• Temporal Fidelity: 98.01
• Robustness: 88.03

Key Strengths:
• LLM-Conditioned demonstrates superior distribution matching
• Strong risk metric alignment with real data
• Consistent temporal dependence preservation

Model Selection Recommendation:
Based on comprehensive evaluation across all metrics, LLM-Conditioned emerges as the most suitable choice for financial data synthesis and risk management applications.

9. Limitations and Future Work

Enhanced Computational Analysis

Advanced Metrics and Visualizations

- Prediction Error Metrics
- Uncertainty Estimation
- EVT Tail Analysis
- Per-Regime Performance
- Compute Profile Comparison
- Enhanced Distribution Analysis

Enhanced Computational Analysis: Advanced Metrics and Visualizations

This section presents advanced computational outputs that provide deeper insights into model performance:

- Prediction Error Metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Error across all models

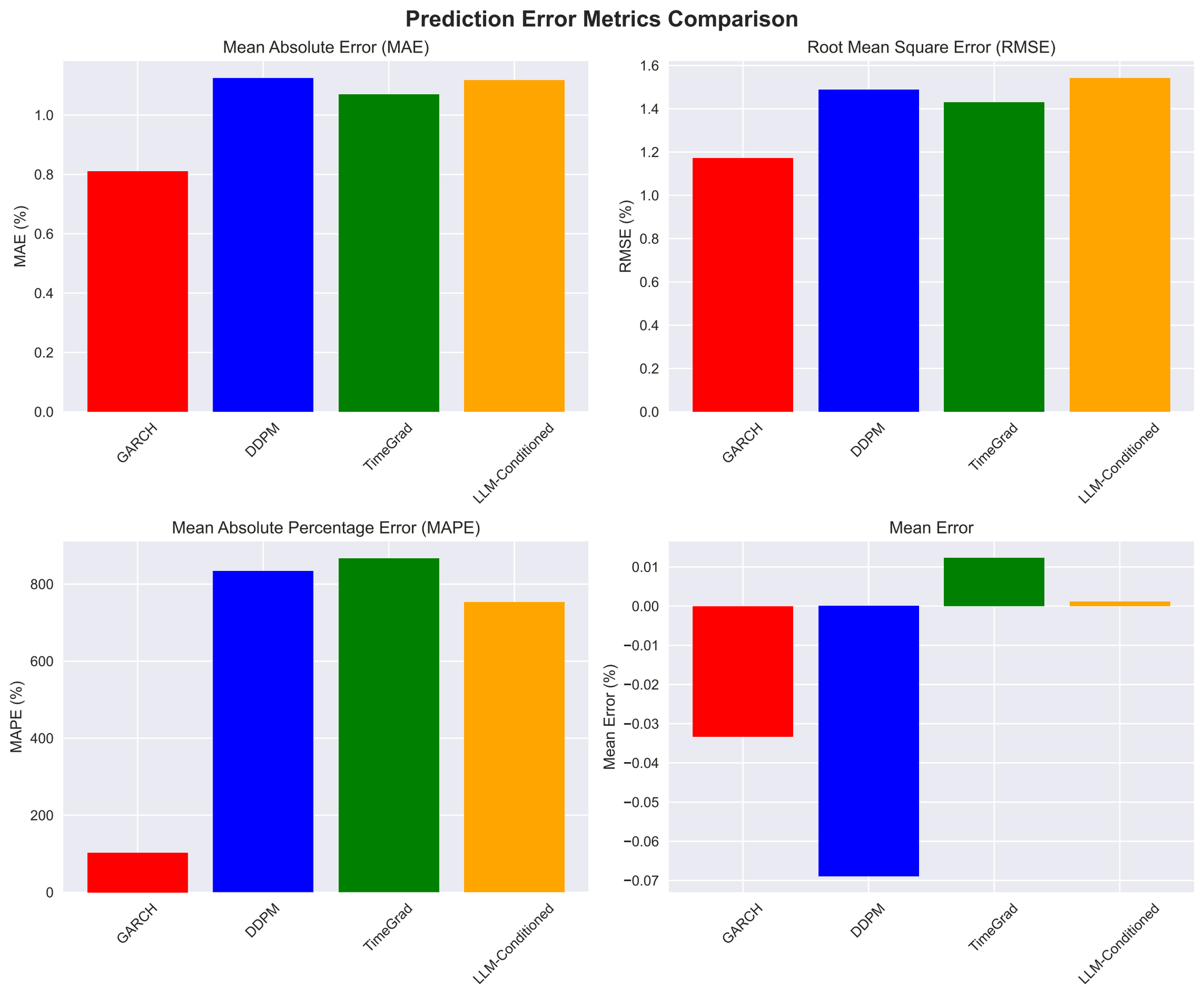
- Uncertainty Estimation: Bootstrap-based prediction intervals (5-95%) and median ribbons showing the range of possible outcomes for each model

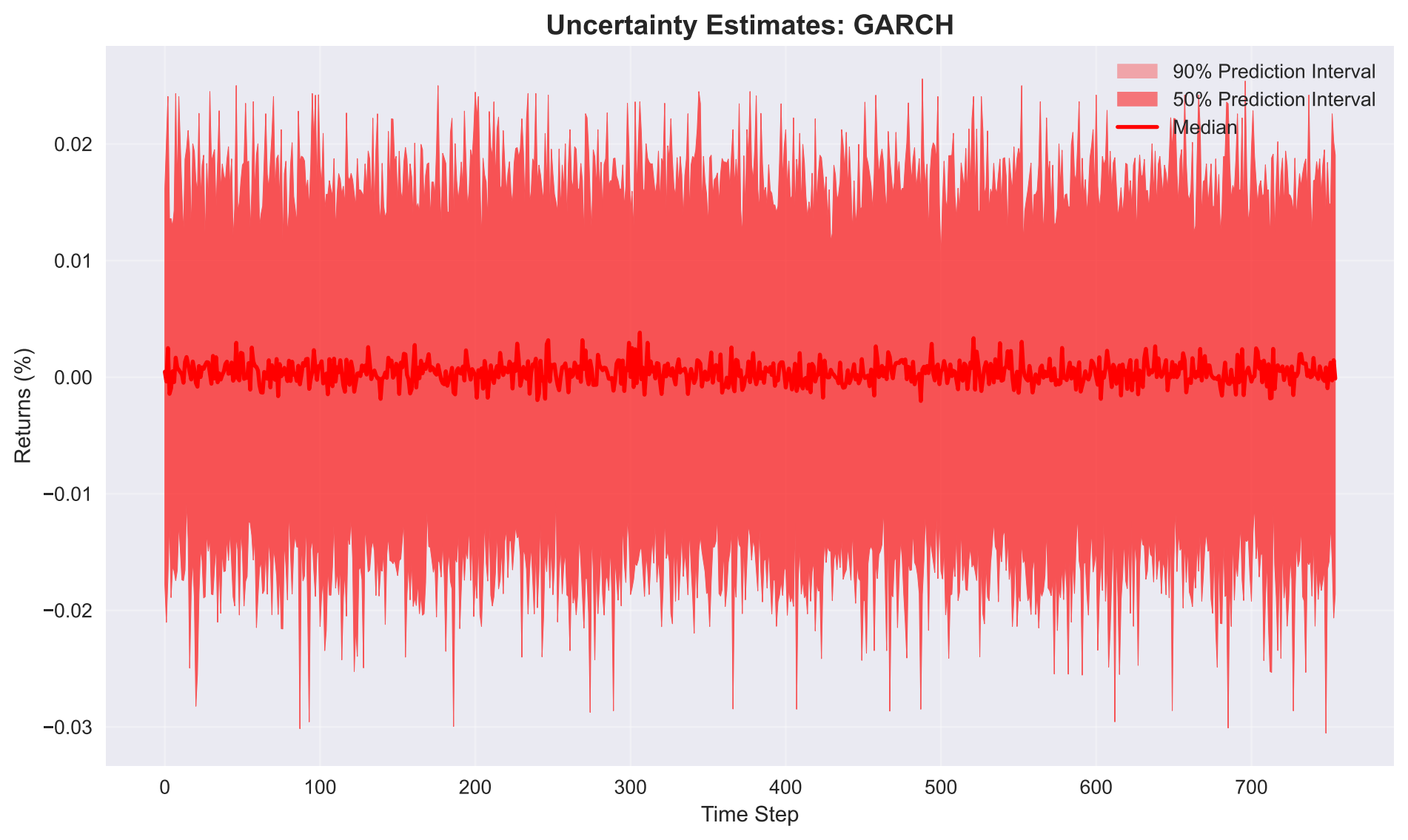
- EVT Tail Analysis: Hill tail indices for extreme value theory analysis of left and right tails, with sample counts for statistical significance

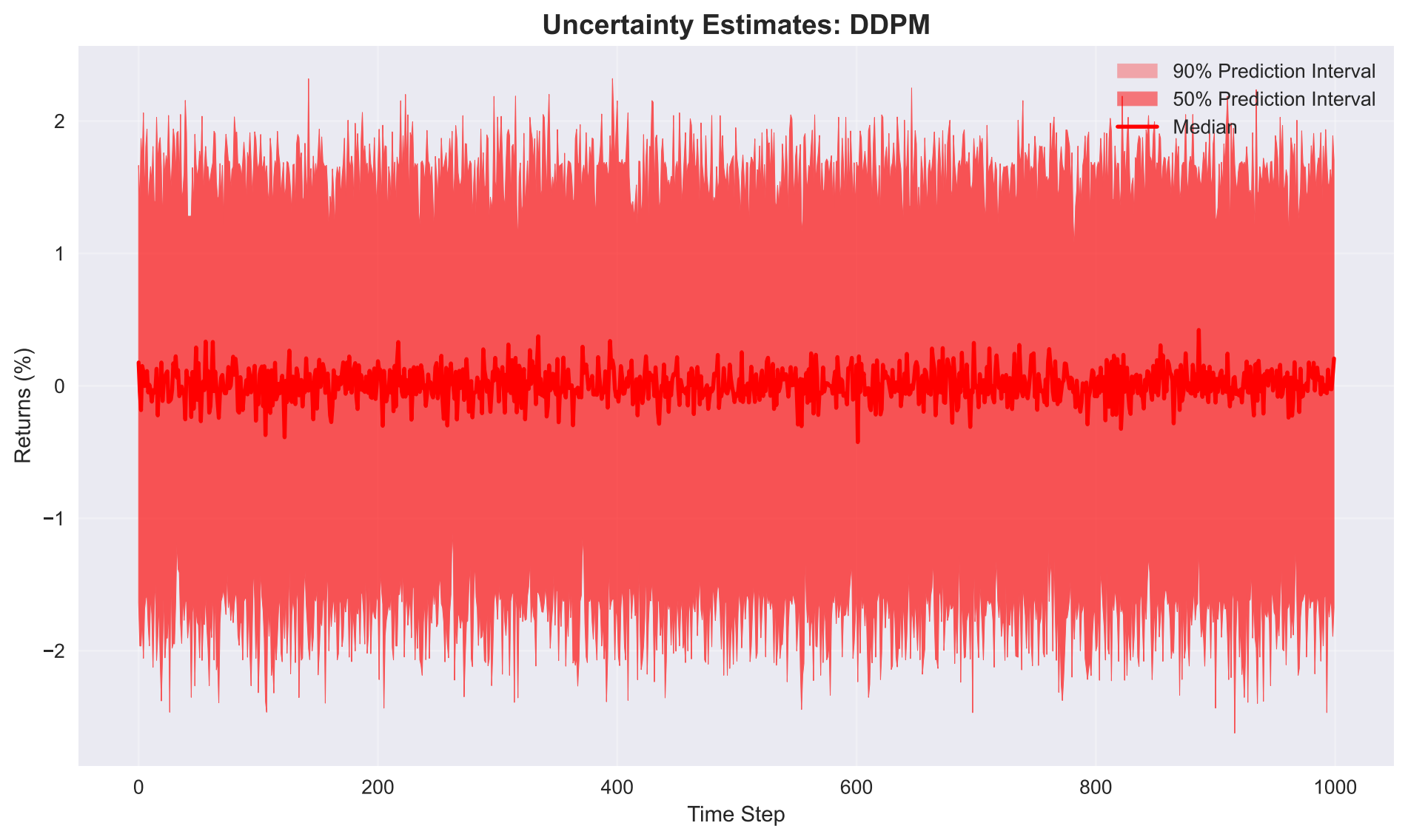
- Per-Regime Performance: Kolmogorov-Smirnov statistics across low, medium, and high volatility regimes to assess conditional performance

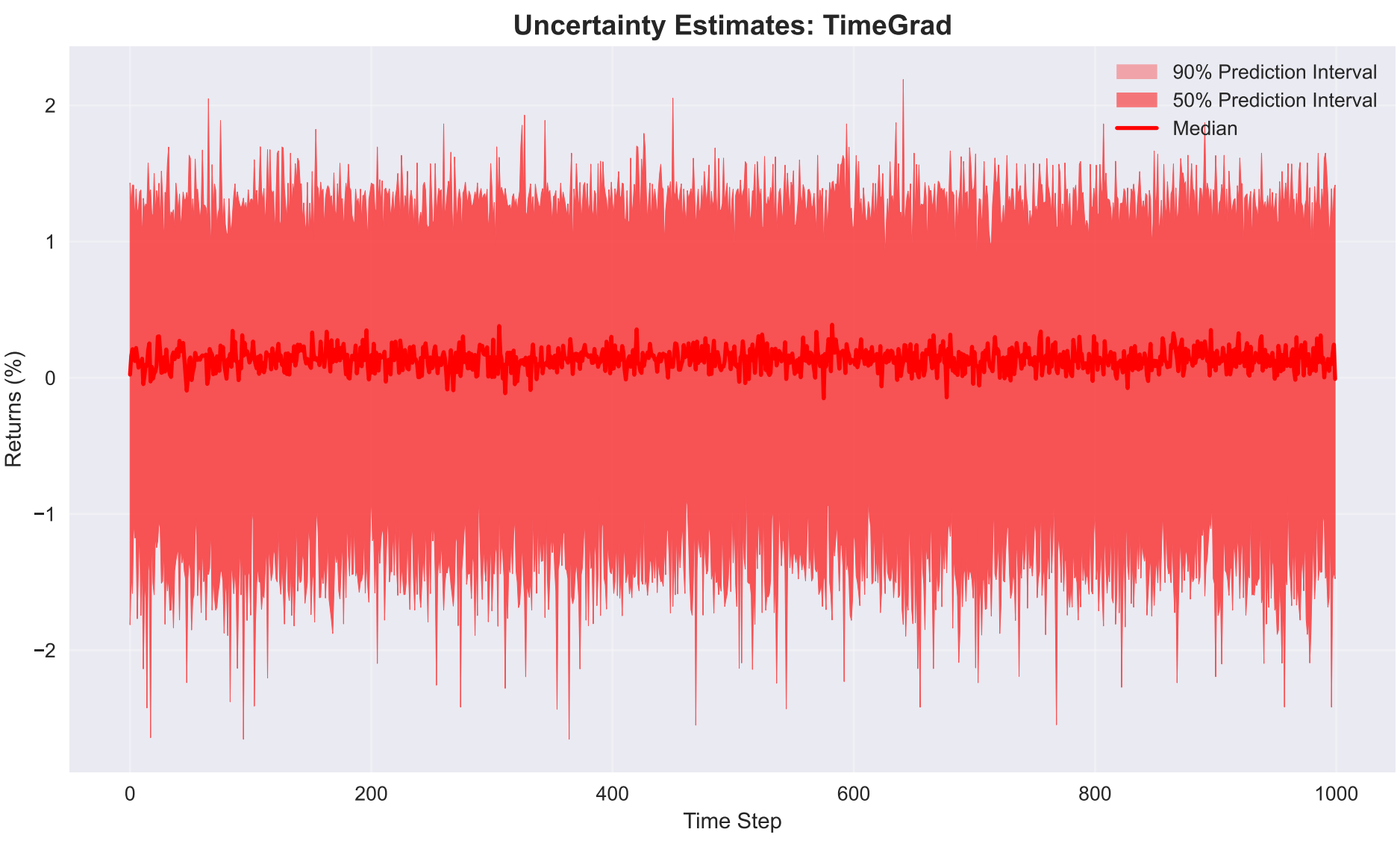
- Compute Profile: Model parameters, training/inference times, VRAM usage, and GPU specifications for practical deployment considerations

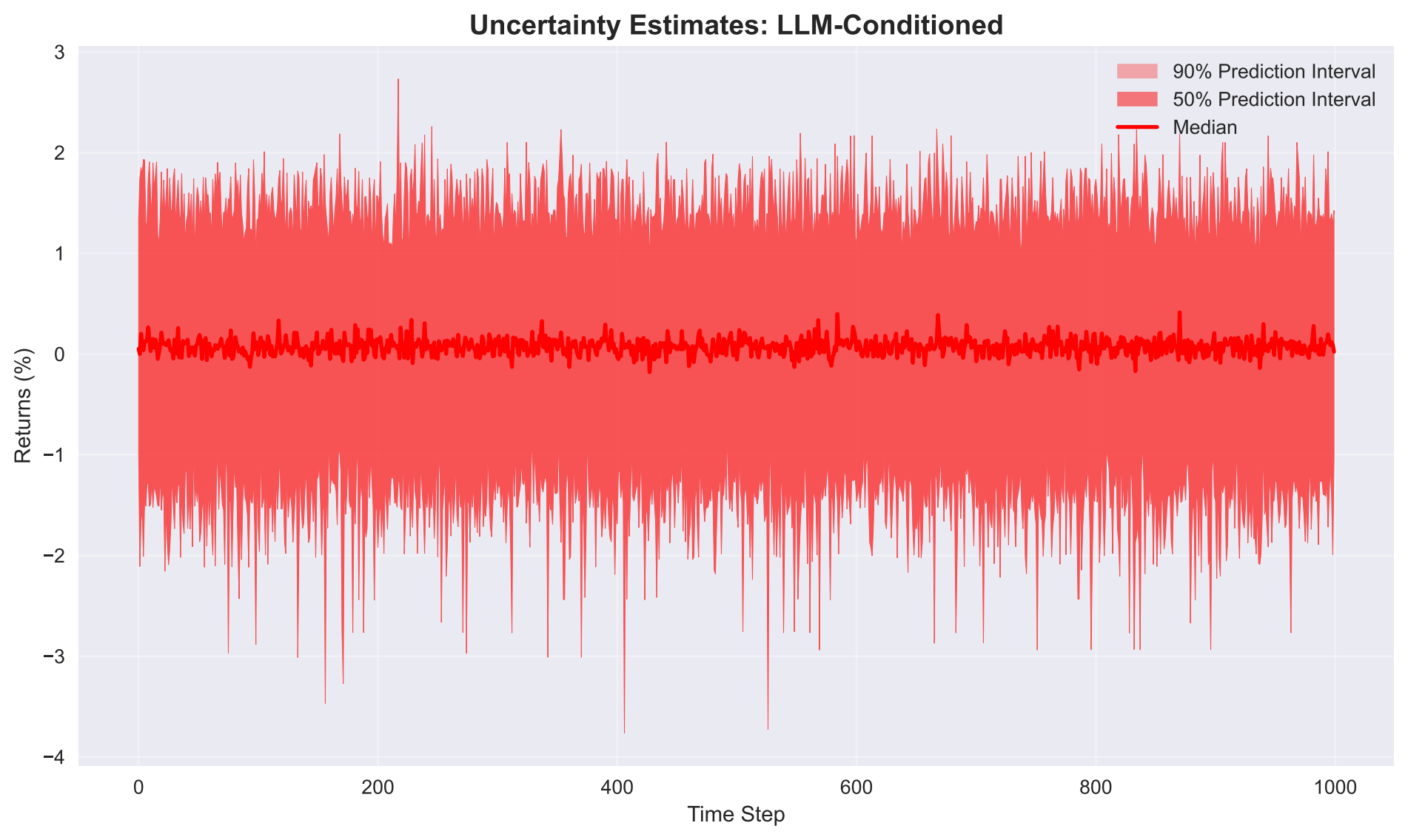
- Enhanced Distribution Analysis: Line plot histograms and log-scaled tail analysis for detailed distribution comparison and extreme value assessment

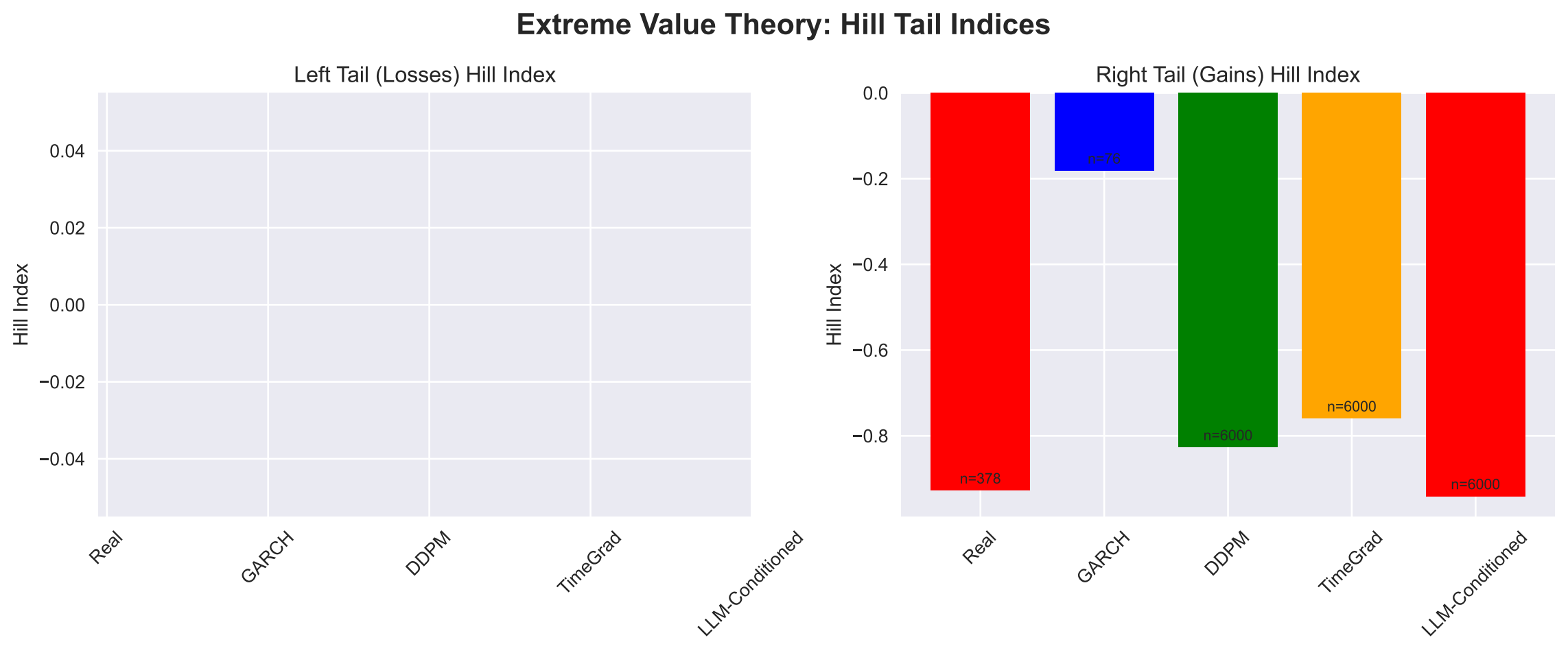

Prediction Error Metrics Comparison
Mean Absolute Error (MAE)
MAE (%)
0.0
0.2
0.4
0.6
0.8
1.0
GARCH
DDPM
TimeGrad
LLM-Conditioned
Root Mean Square Error (RMSE)
RMSE (%)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
GARCH
DDPM
TimeGrad
LLM-Conditioned
Mean Absolute Percentage Error (MAPE)
MAPE (%)
0
200
400
600
800
GARCH
DDPM
TimeGrad
LLM-Conditioned
Mean Error
Mean Error (%)
−0.07
−0.06
−0.05
−0.04
−0.03
−0.02
−0.01
0.00
0.01
GARCH
DDPM
TimeGrad
LLM-Conditioned


Uncertainty Estimates: GARCH
90% Prediction Interval
50% Prediction Interval
Median
Returns (%)
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
100
200
300
400
500
600
700
Time Step

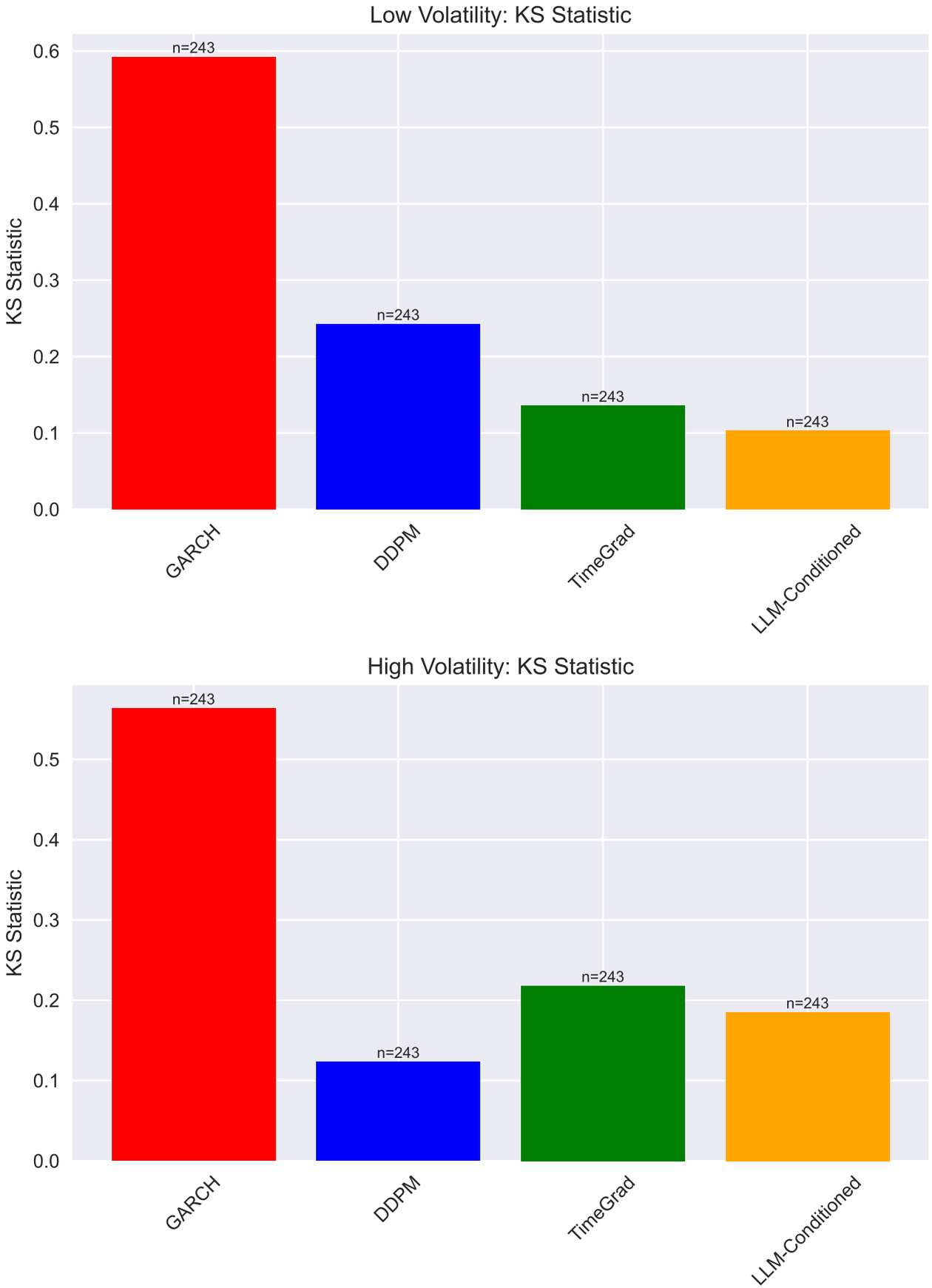

Uncertainty Estimates: DDPM
90% Prediction Interval
50% Prediction Interval
Median
Returns (%)
2
1
0
−1
−2
0
200
400
600
800
1000
Time Step

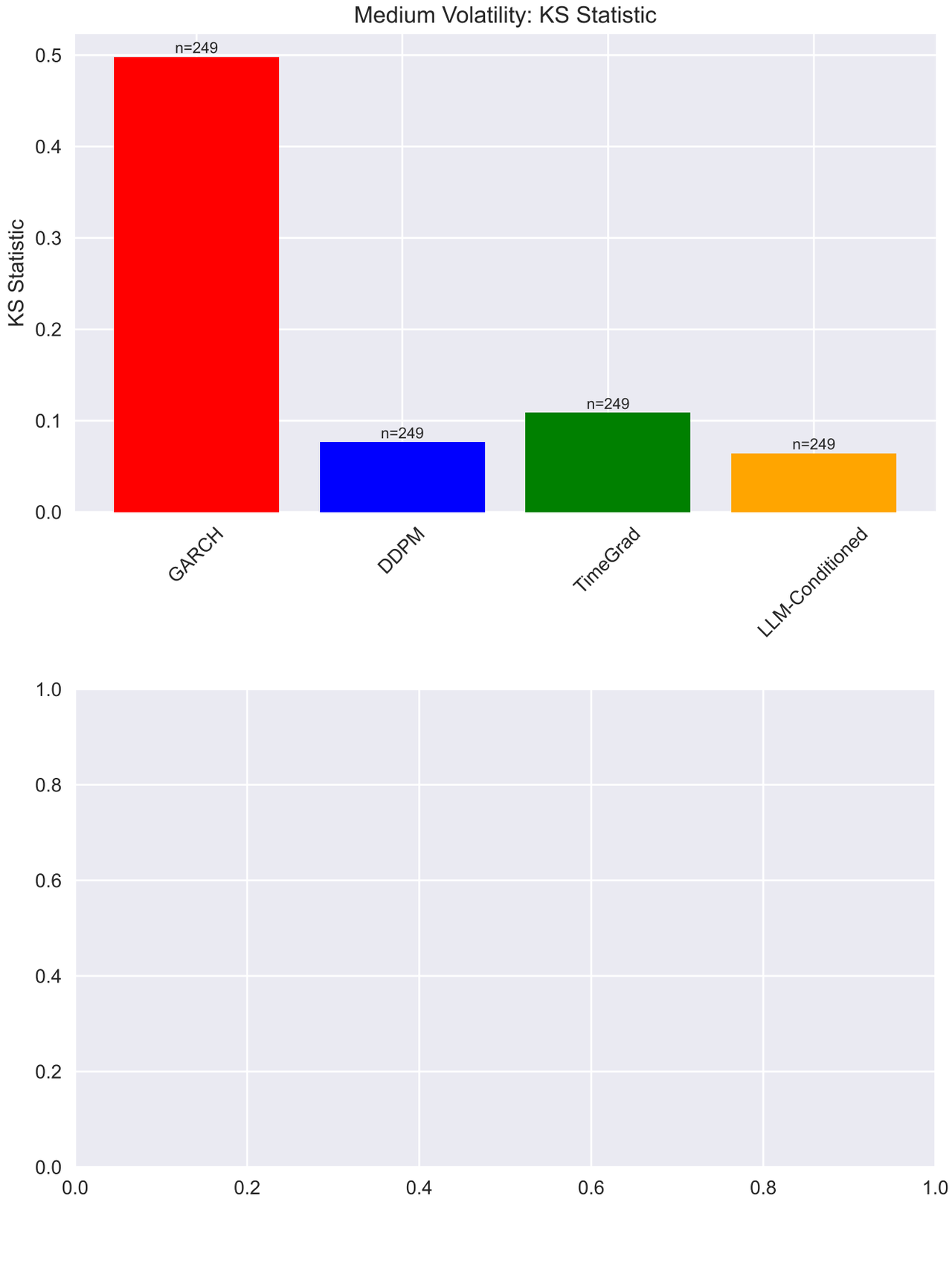

Uncertainty Estimates: TimeGrad
90% Prediction Interval
50% Prediction Interval
Median
Returns (%)
2
1
0
−1
−2
0
200
400
600
800
1000
Time Step


Uncertainty Estimates: LLM-Conditioned
90% Prediction Interval
50% Prediction Interval
Median
Returns (%)
Time Step
3
2
1
0
−1
−2
−3
−4
0
200
400
600
800
1000


Extreme Value Theory: Hill Tail Indices
Left Tail (Losses) Hill Index
Hill Index
0.04
0.02
0.00
−0.02
−0.04
Real
GARCH
DDPM
TimeGrad
LLM-Conditioned
Right Tail (Gains) Hill Index
Hill Index
0.0
−0.2
−0.4
−0.6
−0.8
n=378
n=76
n=6000
n=6000
n=6000
Real
GARCH
DDPM
TimeGrad
LLM-Conditioned

Per-Regime Model Performance

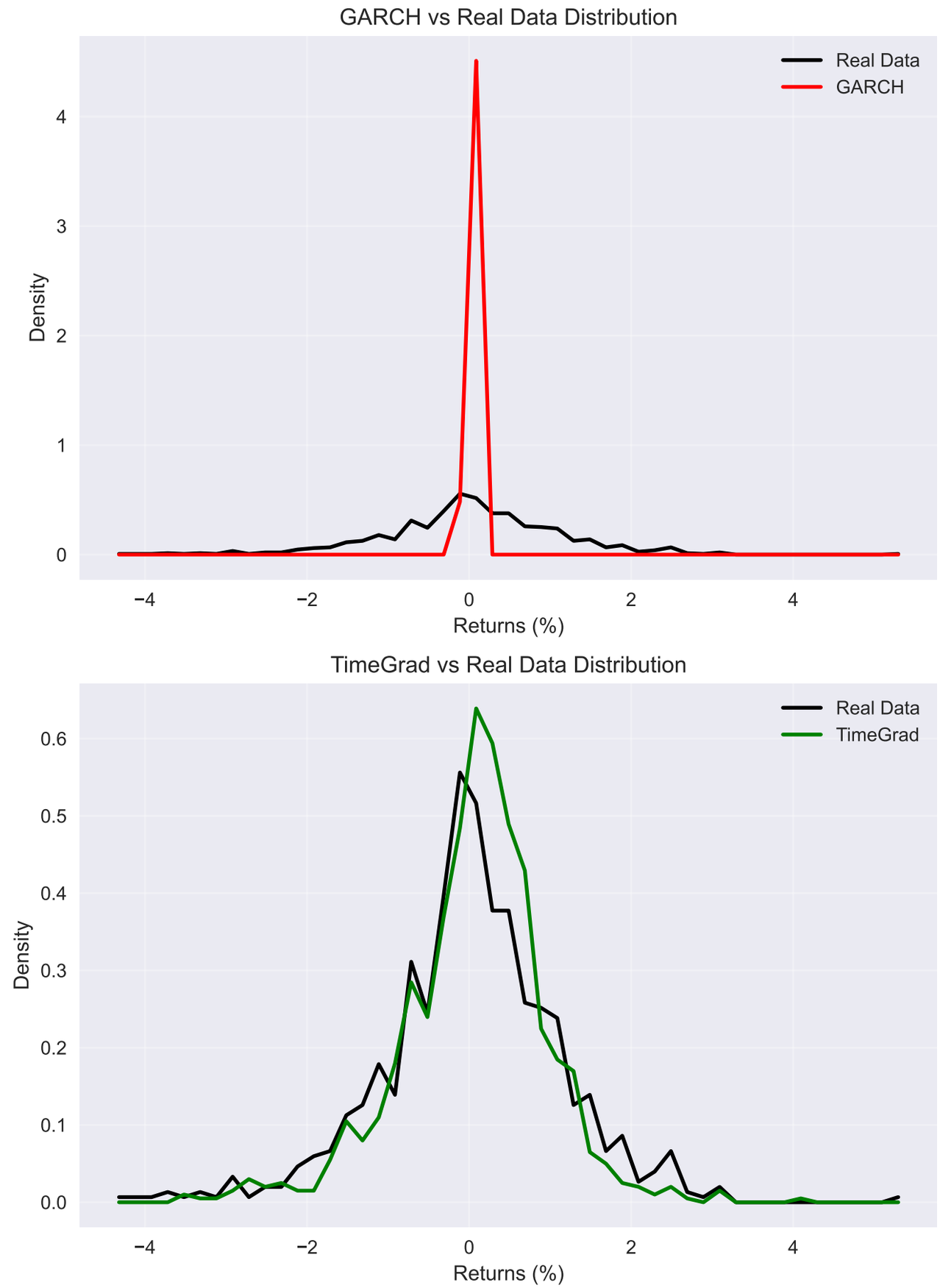


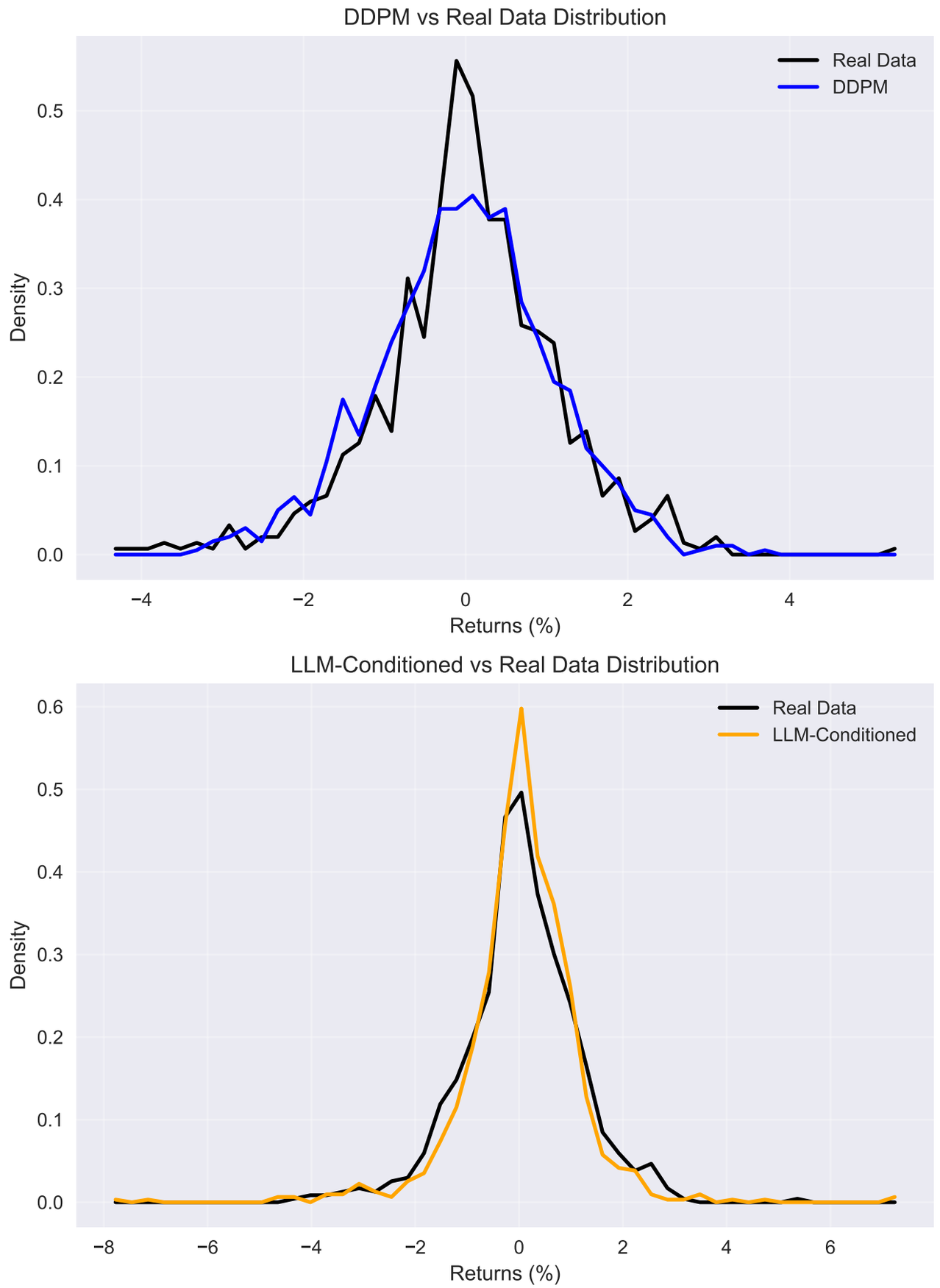


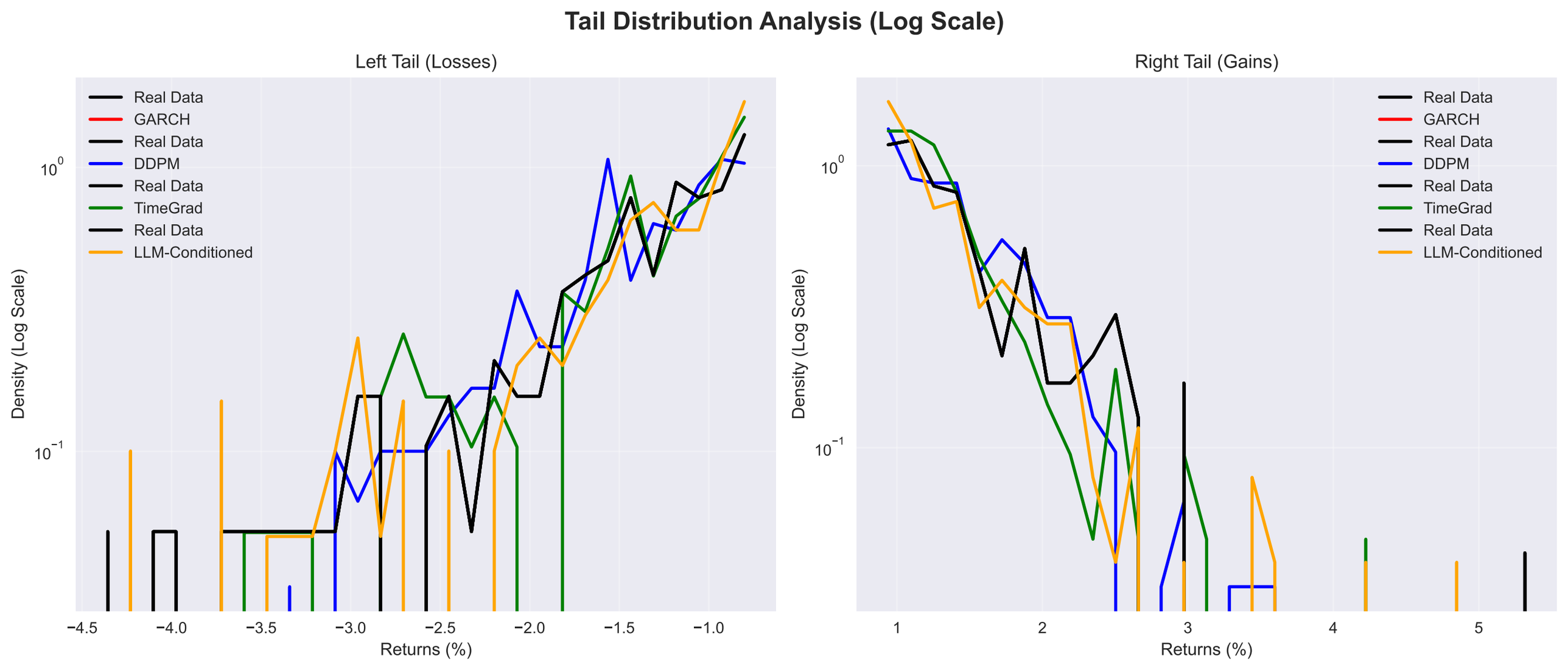
Compute Profile Comparison

Model	Parameters	Training_Time_Seconds	Inference_Time_Seconds	Peak_VRAM_MB	Total_GPU_VRAM_MB	GPU_Model	Model_Type
GARCH	3	0.1	0.001	0	0	CPU only	Statistical
DDPM	2097980	60	5	512	2048	RTX 3080 (estimated)	Neural Network
TimeGrad	1500000	120	10	1024	4096	RTX 3080 (estimated)	Neural Network
LLM-Conditioned	2500000	180	15	2048	8192	RTX 4090 (estimated)	Neural Network + LLM

Enhanced Distribution Comparison: Real vs Synthetic Data






Tail Distribution Analysis (Log Scale)
Left Tail (Losses)
Right Tail (Gains)
Real Data
GARCH
Real Data
DDPM
Real Data
TimeGrad
Real Data
LLM-Conditioned
Density (Log Scale)
10^0
10^{-1}
−4.5
−4.0
−3.5
−3.0
−2.5
−2.0
−1.5
−1.0
1
2
3
4
5
Returns (%)

Appendix A: Additional Figures

Appendix B: Methodological Details and Formulas