








Modelling Financial Markets

Chapter 14: Deep Learning and Time-Series Foundation Models



Daniel Traian PELE

Bucharest University of Economic Studies
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Romanian Academy, Institute for Economic Forecasting
MSCA Digital Finance

Do Deep Networks Forecast Better than Classical Models?



Yann LeCun. Photo: École polytechnique (2025); CC BY-SA 2.0; Wikimedia Commons

"Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction."
— *LeCun, Bengio & Hinton (2015)*

- The signal is in the volatility, not in the mean (S&P 500 since 2000, 6 718 days)
  - Lag-1 autocorrelation: -0.099 for returns, 0.29 for absolute returns
- Returns: no model beats the zero forecast on the S&P 500 and Bitcoin
  - After the correction for multiple testing
- Volatility (SPY, 979 days): mean QLIKE (quasi-likelihood) loss 0.223 for zero-shot Chronos-2 and 0.226 for log-HAR
  - The difference is not significant (Diebold–Mariano test, $p = 0.539$)
- Left tail: raw Chronos-2 breaches its VaR (Value-at-Risk) 1% on 1.67% of S&P 500 days
  - On the S&P 500, filtered historical simulation (FHS) has the lowest loss for the (VaR, ES) pair

▸ Central chart: QLIKE losses of the eight models

Can Deep Networks Forecast Better than the Classical Models?

- **Question**: do deep networks forecast better than the classical models of Chapters 5–9?
  - Targets: **returns**, **volatility** and **tail risk**
  - **Deep learning**: models built from many layers of simple non-linear units, trained on data
  - Time-series **foundation models** (since 2024): networks pretrained on billions of observations from other series, used with no estimation on the target series
  - Why test them: their pretraining corpora are dominated by non-financial series (energy, transport, weather, web)
- Answered with the evaluation tools of Chapter 8
  - Strong benchmarks, loss differentials, backtests
- Part I: networks for sequences, attention, foundation models, Test 1 (returns)
- Part II: Test 2, Test 3, case studies, artificial intelligence (AI) for scientific discovery
  - Test 2: realised volatility
  - Test 3: left-tail risk (value-at-risk, VaR 1%, and expected shortfall, ES 2.5%)

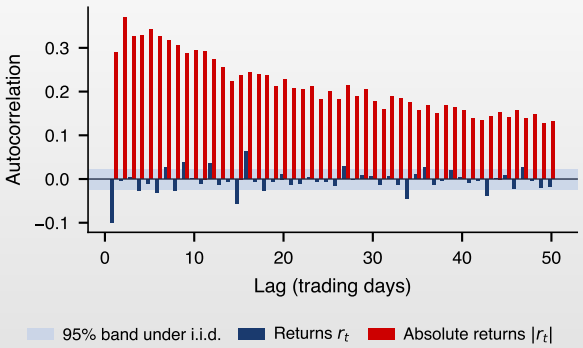
What You Should Be Able to Do

- Derive vanishing gradients from the impulse responses of a vector autoregression (VAR)
- Derive the cell state of the LSTM as a first-order autoregression, AR(1)
  - LSTM: long short-term memory, a gated recurrent network
- Compute point, volatility and VaR/ES forecasts from a zero-shot quantile grid
  - Zero-shot: produced without any training on the target series
- Interpret the comparison with the right benchmark for each target
  - The zero forecast for returns
  - Log-HAR for volatility
  - FHS for the tails
- Interpret the tests of a new model: nesting, multiplicity, calibration, contamination and power
- Detailed examples and advanced topics are in Further Topics, after the Conclusions
  - the footer button leads there; the Back button returns to the starting slide
- Further Topics and Appendix: reading for familiarity

Three Forecasting Targets

- **Returns** $r_t = 100 \ln(P_t / P_{t-1})$ (conditional mean), P_t the close on day t
  - Very low signal-to-noise ratio: predictable patterns are traded away (Chapter 2)
  - Benchmark: the zero forecast
- **Volatility** (conditional variance): persistent and clustered (Chapters 1 and 5), a signal to learn
  - Benchmark: log-HAR, log realised variance regressed on its daily, weekly and monthly averages (heterogeneous autoregressive, Test 2)
- **Tail risk**: the shape of the lower tail, not only the centre (Chapter 7)
  - VaR 1% (value-at-risk): the daily loss exceeded with probability 1%
  - ES 2.5% (expected shortfall): the average loss on the worst 2.5% of days
  - Benchmark: FHS (filtered historical simulation)
  - FHS: the empirical quantile of returns standardised by a GARCH volatility (generalised autoregressive conditional heteroskedasticity, Chapter 5), times the forecast volatility

Where Is the Signal?



- S&P 500 daily returns since 2000 (6 718 days), lags 1–50
  - Light-blue band: $\pm 1.96/\sqrt{T} = \pm 0.024$ (T: number of days), the 95% range of an autocorrelation under i.i.d. returns (independent and identically distributed)
- Returns: lag-1 autocorrelation -0.099
  - 15 of 50 lags outside the band, small and of changing sign
- Absolute returns: 0.29 at lag 1, 0.13 at lag 50
  - All 50 lags outside the band: persistent volatility clustering
  - The band assumes $|r_t|$ i.i.d. with $E[r_t^2] < \infty$; otherwise use a robust band
- Flexible models can find the volatility signal, not much in the mean

MFM_ch14_acf_signal ▶ Interactive

From the Multilayer Perceptron to Deep Learning (1/3)

- **Question of the section**: how can a network use information from distant days?
  - And why does a simple recurrent network fail?
- MLP (multilayer perceptron, Chapter 13)

$$h^{(\ell)} = \varphi\left(W^{(\ell)} h^{(\ell-\mathbf{1})} + b^{(\ell)}\right), \qquad h^{(\mathbf{0})} = x$$

  - Each layer applies a non-linear function to a linear combination of the previous layer
  - x: input vector; $h^{(\ell)}$: output of layer ℓ
  - $W^{(\ell)}$, $b^{(\ell)}$: weight matrix and bias vector of layer ℓ, estimated from the data
  - φ: activation, applied element by element: ReLU (rectified linear unit) $\max(0, z)$, tanh, or the sigmoid $\sigma(z) = 1/(1 + e^{-z}) \in (0, 1)$
- An MLP takes a fixed window of inputs
  - A sequence network must carry information from one step to the next

From the Multilayer Perceptron to Deep Learning (2/3)

- Training: backpropagation (Rumelhart, Hinton & Williams (1986)) computes the gradient of the loss with respect to every weight
- Update rule: Adam (adaptive moment estimation, Kingma & Ba (2015))
  - A gradient step normalised by exponential averages of past gradients and of their squares
- Loss for point forecasts: the mean squared error (MSE), minimised by the conditional mean
- Loss for a quantile of level u: the pinball loss

$$\rho_u(y, q) = (u - 1\{y < q\})(y - q)$$

  - y: realised value; q: quantile forecast
  - Asymmetric: values above q cost u per unit, values below cost $1 - u$
  - The expected loss is minimised by the true u-quantile

From the Multilayer Perceptron to Deep Learning (3/3)

- Universal approximation (Hornik, Stinchcombe & White (1989))
  - A wide enough network approximates any continuous function
  - Silent on the sample size needed, the binding constraint with one financial series
  - Regularisation: early stopping, weight decay (a penalty on the size of the weights), averages over random initialisations (seeds)
  - The validation block must follow the training block
- 2024 Nobel Prize in Physics to Hopfield and Hinton
  - The methods: LeCun, Bengio & Hinton (2015)
  - In finance, Gu, Kelly & Xiu (2020) gain on a cross-section of thousands of stocks, not on one index
  - Fischer & Krauss (2018) apply LSTMs to S&P 500 constituents, forecasting which stocks beat the next-day cross-sectional median



Geoffrey Hinton, Nobel laureate 2024

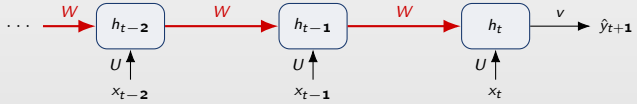
Photo: Arthur Petron (2024); CC BY-SA 4.0; Wikimedia Commons

RNN and LSTM (1/2)

- **Recurrent neural network** (RNN): a hidden state is updated at every step

$$h_t = \tanh\left(W h_{t-1} + U x_t + b\right), \qquad \hat{y}_{t+1} = v^\top h_t + c$$

  - The update uses its previous value and the new input
  - h_t: hidden state, the memory of the network at t; x_t: input at t (e.g. today's return); $\hat{y}_{t+1}$: forecast
  - W, U, b: recurrent weights, input weights and bias; v, c: output weights and intercept; tanh keeps h_t in $(-1, 1)$
  - The same weights at every step: unrolled over time, a very deep network with shared weights



RNN and LSTM (2/2)

- **Long short-term memory** (LSTM, Hochreiter & Schmidhuber (1997)): gates decide what to forget, what to write and what to show

$$f_t, i_t, o_t = \sigma(W_\bullet x_t + U_\bullet h_{t-1} + b_\bullet), \quad g_t = \tanh(W_g x_t + U_g h_{t-1} + b_g)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \quad h_t = o_t \odot \tanh(c_t)$$

- Notation
  - $W_\bullet$ weights the input and $U_\bullet$ the previous state here, reversing the RNN letters
  - f_t: forget gate, i_t: input gate, o_t: output gate; each has its own weights $W_\bullet$, $U_\bullet$, $b_\bullet$, $\bullet \in \{f, i, o\}$
  - σ: sigmoid; a gate near 0 blocks, near 1 lets through
  - g_t: candidate new content; c_t: cell state, the long-term memory
  - h_t: output; $\odot$: element-by-element product
- GRU (gated recurrent unit): a lighter variant with two gates

Vanishing Gradients: a Bound and a VAR(1) (1/2)

- Training needs the effect of old states on today's state
- Backpropagation through time multiplies one Jacobian per step

$$\frac{\partial h_T}{\partial h_{T-k}} = \prod_{j=0}^{k-1} D_{T-j}\, W, \qquad D_t = \operatorname{diag}\left(\varphi'(W h_{t-1} + U x_t + b)\right)$$

  - $\partial h_T / \partial h_{T-k}$: the matrix of changes in h_T per unit change in the state k steps earlier
  - D_t: diagonal matrix of the slopes φ' of the activation at step t
- Bound, by submultiplicativity $\|AB\|_2 \le \|A\|_2 \|B\|_2$ (Pascanu, Mikolov & Bengio (2013))

$$\left\| \frac{\partial h_T}{\partial h_{T-k}} \right\|_2 \le \left(\gamma_\varphi \|W\|_2\right)^k, \qquad \gamma_\varphi = \sup_z |\varphi'(z)|$$

  - $\|W\|_2$: spectral norm, the largest factor by which W stretches a vector
  - γ_φ: the largest slope of the activation ($= 1$ for tanh)
  - $\gamma_\varphi \|W\|_2 < 1$: the gradient shrinks geometrically with k, i.e. it vanishes
- Scalar case: with $|w| = 0.9$ the gradient falls to 0.12 after 20 steps and to 0.005 after 50 steps (Bengio, Simard & Frasconi (1994))
  - Remedies: clipping the gradient norm (explosion), gating as in the LSTM (vanishing)

Vanishing Gradients: a Bound and a VAR(1) (2/2)

- Linearise around a fixed point h^* of the network

$$h_t - h^* \approx A\,(h_{t-1} - h^*) + DUx_t, \qquad A = DW$$

  - Fixed point: a state that reproduces itself when the input is zero
  - D: the slopes of the activation at h^*
  - The hidden state follows a VAR(1) (vector autoregression) with coefficient matrix A
- $\partial h_T / \partial h_{T-k} = A^k$ is the impulse response of this VAR at horizon k
  - $\rho(A)$: spectral radius, the largest absolute eigenvalue of A
  - $\rho(A) < 1$: stable VAR, $\|A^k\| \to 0$ geometrically as $k \to \infty$
  - At finite horizons $\|A^k\|$ can grow transiently if A is not normal ($AA^\top \neq A^\top A$)
  - $\rho(A) > 1$: they explode
- A stable RNN near its fixed point is a short-memory model, like any stationary VAR
  - Seminar 14, exercise A4
  - Details in the Appendix

LSTM Memory: the Forget Gate

- Along the cell state the only multiplier is the forget gate
  - $\partial c_T / \partial c_{T-k} = \prod_{j=0}^{k-1} f_{T-j}$ on the direct path
  - $c_t = f_t \odot c_{t-1} + i_t \odot g_t$: an AR(1) with a random coefficient
- Simplified scalar cell with a constant forget gate f: an AR(1) with coefficient f
  - $f = \sigma(b_f)$ when the gate inputs are zero; b_f: its bias
  - With a white-noise input $i_t g_t$ and a stationary start, $\mathrm{Corr}(c_t, c_{t-k}) = f^k$ decays geometrically
  - With $b_f = 0$, the forget gate is $f = 0.5$, and after 50 steps $0.5^{50} \approx 9 \times 10^{-16}$ of the memory is left
  - $b_f = 3$: $f = 0.953$, and $0.953^{50} \approx 0.09$ of the memory is left after 50 steps
  - Half-life 14 steps (the k with $f^k = \frac{1}{2}$, $k = \ln 0.5 / \ln f$)
- Long memory decays hyperbolically: $\rho_k \sim C k^{2d-1}$
  - With ρ_k the autocorrelation at lag k
  - $d \in (0, \frac{1}{2})$ the memory parameter, $C > 0$ a constant
  - S&P 500 absolute returns: the fitted geometric decay gives $\rho_{150} = 0.027$, against 0.059 observed (▸ Appendix)



Sepp Hochreiter, co-inventor of the LSTM

Photo: Eulenreich (2015); CC BY-SA 4.0; Wikimedia Commons

Attention: Queries, Keys and Values (1/2)

- **Question of the section**: how can a network link today directly to a day 500 days back?
  - Without the product of Jacobians that makes the RNN gradient vanish
- Each time step asks a question (query) and compares it with the key of every step it is allowed to see

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \qquad Q = XW_Q,\ K = XW_K,\ V = XW_V$$

  - The answer is a weighted average of their values (Vaswani et al. (2017))
  - X: the inputs, one row per time step
  - W_Q, W_K, W_V: learned weight matrices
  - Q, K, V: queries, keys and values

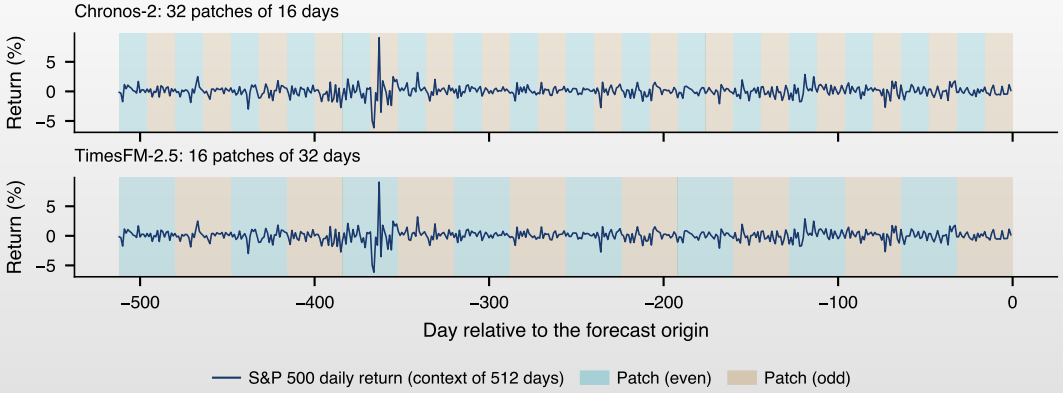
Attention: Queries, Keys and Values (2/2)

- $QK^\top$: the similarity score of every pair of steps
- d_k: length of a key vector
  - Dividing by $\sqrt{d_k}$ keeps the scores from saturating the softmax
- softmax turns each row of scores $s_1, \ldots, s_n$ into positive weights that sum to one

$$\text{softmax}(s)_j = \frac{e^{s_j}}{\sum_{i=1}^{n} e^{s_i}}$$

- No product of Jacobians over time: any past step is one hop away
- Cost $O(n^2)$: the work grows with the square of the number of inputs n

Patching a 512-Day Context



MFM_ch14_tokenization ▶ Interactive

Positions and Patches

- **Positional encoding**: attention treats its inputs as a set, so order is added
  - Sinusoids in the original Transformer; rotary embeddings in Chronos-2
  - Rotary embeddings (RoPE): each query and key is rotated by an angle proportional to its position
  - So the score of two steps depends on how far apart they are, not on where they sit in the context
- **Patching** (Nie et al. (2023)): a token is a block of consecutive observations
  - A learned map turns it into a vector of fixed length (embedding)
  - 512 daily returns in patches of 16 days: 32 tokens, 256 times fewer attention scores (32^2 instead of 512^2)
- Chart: the last 512 daily returns of the S&P 500
  - Split as Chronos-2 (16 days) and TimesFM-2.5 (32 days) read them
  - Each shaded block becomes one token
- Global models learn across many series: DeepAR (Salinas et al. (2020)), N-BEATS (Oreshkin et al. (2020)), PatchTST (Nie et al. (2023))
  - Foundation models push this idea to very large corpora

MFM_ch14_tokenization

What Is a Foundation Model?

- **Question of the section**: can a model pretrained on other series forecast a financial series it never saw in training?
- Foundation model: a large network pretrained once on a very broad corpus, then used on new tasks
  - An idea borrowed from large language models (LLM)
- Appeal for finance: one model for thousands of series, no estimation per series
- Three families of models
  - **Chronos** (Ansari et al. (2024)): values become the tokens of a language model
  - **TimesFM** (Das et al. (2024)): a decoder-only Transformer that reads the series in patches
  - **Moirai** (Woo et al. (2024)): any number of variables and a mixture-distribution output (▸ Appendix)
  - Moirai is not tested here because of its non-commercial licence

Zero-Shot Forecasting: What Goes In and What Comes Out

- **Zero-shot** forecasting: the parameters $\hat{\theta}$ are fixed in pretraining

$$\left(\hat{q}_{u_1, T+h}, \ldots, \hat{q}_{u_K, T+h}\right)_{h=1,\ldots,H} = F_{\hat{\theta}}\left(y_{T-C+1}, \ldots, y_T\right)$$

  - At time T the model receives only the recent history of the series (the **context**)
  - $F_{\hat{\theta}}$: the pretrained network, the same for every series
  - C: context length (512 days here); H: forecast horizon
  - $\hat{q}_{u_k, T+h}$: the u_k-quantile of the value at $T+h$, so the output is a quantile grid, not only a point
- Fine-tuning is possible but not used
  - Fine-tuning: further training on the target series
  - We test the "no estimation" promise itself
- The specific risk: the target series may be part of the pretraining corpus (slide on leakage)

Chronos: Time Series as a Language

- Ansari et al. (2024): turn values into tokens and train a language model with cross-entropy
  - The language model: T5, Text-to-Text Transfer Transformer
- Why tokens: the language-model architecture stays unchanged
  - A distribution over tokens imposes no parametric shape (it can be skewed or multimodal)
- Mean scaling and quantisation

$$\tilde{x}_t = \frac{x_t}{s}, \quad s = \frac{1}{C}\sum_{i=\mathbf{1}}^{C} |x_i|, \qquad z_t = \arg\min_j |\tilde{x}_t - c_j|, \quad c_j = -15 + \frac{30\,(j-1)}{4{,}092}$$

  - s: mean absolute value of the C values in the context
  - c_j, $j = 1, \ldots, 4{,}093$: equally spaced bin centres in $[-15, 15]$, $30/4{,}092 = 0.0073$ apart, plus special tokens
  - z_t: the token of x_t, the index of the nearest centre
  - Forecast: sample future tokens z, map them back to values $s\,c_z$, read quantiles from the samples
- Pretraining corpus: public datasets (energy, transport, retail, weather, web, finance and economics)
  - Plus synthetic TSMixup series: random convex combinations of real series
  - And KernelSynth series: paths of Gaussian processes with randomly combined covariance kernels
- Chronos (T5): 8 to 710 million parameters
  - Chronos-Bolt, tested here: 9 to 205 million

The Three Models We Test

Model	Parameters	Patch	Quantile levels	Weights released
Chronos-Bolt (small)	48 M	16	10%–90%	25.11.2024
TimesFM-2.5	231 M	32	mean, 10%–90%	01.10.2025
Chronos-2	119 M	16	21: 1%, 5%, 10%, . . . , 95%, 99%	30.10.2025

- **Chronos-Bolt**: patches and a direct quantile output instead of token sampling
- **Chronos-2** (Ansari et al. (2025)): robust scaling, attention across time and across groups of related series
- **TimesFM** (Das et al. (2024)): pretrained on about 10^{11} time points
- Parameters are counted on the loaded models
  - TimesFM-2.5 is released as timesfm-2.5-200m
- All run on a CPU (Central Processing Unit) with a 512-day context
  - The release dates of the weights bound what the models could have seen
  - Other models and their corpora in the Appendix

From a Quantile Grid to VaR and ES (1/2)

- The model returns quantiles q_u at levels $u_1 < \cdots < u_K$: $P(Y \leq q_u) = u$
  - Y: the future value; K: the number of levels
- Why a grid: the same quantiles give the point forecast, the volatility and the tail risk
  - So one model serves all three tests
- Point forecasts from the grid
  - the median $q_{0.5}$, or the mean $\int_0^1 q_u \, du$, the average of all quantiles
  - for any function g with a finite mean: $E[g(Y)] = \int_0^1 g(q_u) \, du$, because Y has the same distribution as q_U with U uniform (e.g. $g = \exp$ when Y is a log variance)
  - If g is increasing, $g(q_u)$ is also the u-quantile of $g(Y)$
  - Check: Y uniform on $(-1, 1)$, $q_u = 2u - 1$, $g(y) = y^2$: $\int_0^1 (2u-1)^2 \, du = 1/3 = E[Y^2]$, although $(2u-1)^2$ is not a quantile of Y^2 (it falls, then rises in u)

From a Quantile Grid to VaR and ES (2/2)

- Tail risk at level α (Chapter 7), for a return Y

$$\mathrm{VaR}_\alpha = -q_\alpha, \qquad \mathrm{ES}_\alpha = -\frac{1}{\alpha}\int_0^\alpha q_u \, du$$

- Below the lowest level the grid is silent
  - Holding q_u at $q_{0.01}$ understates ES
  - Interpolation can err either way (Seminar 14, exercises A1 and A7)
- Chronos-Bolt and TimesFM stop at 10%: a "VaR 1%" request returns the 10% quantile
- The chapter's rule: linear interpolation between levels
  - q_u held flat below the lowest level
  - The worked example is in Further Topics

Pretraining Leakage and Contamination

- A test is out of sample only if the test period was never seen during training
- It can be "seen" in four ways
  - (i) the same series is in the pretraining corpus
  - (ii) correlated series carry the same shocks
  - (iii) developers select model versions on overlapping public benchmarks
  - (iv) our own look-ahead: scalings, windows, seeds or contexts chosen with test data
- The Chronos corpus includes financial series (e.g. daily exchange rates 1990–2016)
  - So we cannot rule out that our series were seen in pretraining
- Remedy for (i)–(iii): a **post-release window** from 3 November 2025
  - A common cutoff after all three sets of weights were published
  - Short: about 220 trading days (320 calendar days for Bitcoin), hence low power, especially for VaR 1%
  - Free of weight contamination only if the checkpoint used predates the window: its revision and weight hash are pinned in the code (appendix on the corpora)
- (iv) is under our control: every scaling, window and estimate uses only past data
  - Choosing the context or the hyperparameters with hindsight remains possible after the release
  - Formal protocol in the Appendix

Design of the Return Test

- **Question**: does any model forecast tomorrow's daily return better than the zero forecast?
- Test period: S&P 500 and BET from 04.01.2016, Bitcoin from 01.01.2019, to 18.09.2026
  - BET: Bucharest Stock Exchange index; 2 693 days (S&P 500), 2 675 (BET) and 2 818 (Bitcoin)
  - About ten years (eight for Bitcoin), including the March 2020 crash and 2022
  - Estimation also uses earlier data: the series start in 1990 (S&P 500), 1997 (BET) and 2014 (Bitcoin)
- Why three markets
  - S&P 500: the most liquid equity market, where a simple signal would already be traded away
  - BET: a smaller, less liquid market, where stale prices can create predictability
  - Bitcoin: traded around the clock, with a daily standard deviation almost three times larger (3.21% against 1.13% for the S&P 500)
- Models: historical mean, AR(1), LSTM, Chronos-2, Chronos-Bolt and TimesFM-2.5
  - AR(1): $\hat{r}_t = \hat{a} + \hat{b}\, r_{t-1}$, least squares on the last 1,000 days, re-estimated daily
  - Foundation models: the grid median, a point forecast that extreme values do not move
  - Grid means are reported separately
- Every forecast uses only data up to the previous close

Judging the Return Forecasts (1/2)

- Benchmark: the zero forecast; out-of-sample

$$R^2 = 1 - \frac{\sum_t (r_t - \hat{r}_t)^2}{\sum_t r_t^2}$$

  - r_t: realised return; $\hat{r}_t$: forecast from data up to $t-1$
  - $R^2 > 0$: smaller squared errors than the zero forecast
- Why zero rather than the historical mean: the estimated mean adds estimation noise
  - Zero is the benchmark of Gu, Kelly & Xiu (2020), the historical mean that of Campbell & Thompson (2008)
- Grid means ($\int_0^1 q_u \, du$) do even worse than the medians
  - On the S&P 500, R^2 is -0.77% (Chronos-2) and -3.42% (TimesFM-2.5)

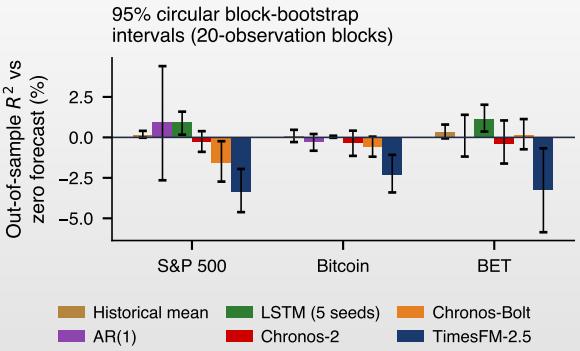
Judging the Return Forecasts (2/2)

- Block-bootstrap confidence interval (CI) for R^2
  - Rebuild a test period of the same length from blocks of 20 consecutive days drawn with replacement, 1,000 times
  - Keep the 2.5% and 97.5% percentiles
  - The blocks keep volatility clusters together
- Diebold–Mariano (DM, Diebold & Mariano (1995)) test on squared errors

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{\hat{V}_{NW}/n}}, \qquad d_t = (r_t - \hat{r}_t)^2 - r_t^2$$

  - $\bar{d}$: mean of d_t over the n forecasts
  - $\mathrm{DM} > 0$: the model has larger errors than the zero forecast
  - $\hat{V}_{NW}$: long-run variance of d_t, estimated by Newey–West, a HAC (heteroskedasticity and autocorrelation consistent) estimator
  - Under H_0: $E[d_t] = 0$, DM is approximately standard Normal
  - The approximation needs a finite, non-zero long-run variance of d_t and weak dependence
  - Here $d_t = \hat{r}_t^2 - 2r_t\hat{r}_t$: the moments of $r_t\hat{r}_t$ matter, not necessarily $E[r_t^4]$

Out-of-Sample R^2 for Returns



- R^2 against the zero forecast (%), 95% block-bootstrap CIs
  - Full table with DM p-values in the ▸ Appendix
- S&P 500 and Bitcoin: no model beats zero in the two-sided DM test
- On the S&P 500, the historical mean passes the Clark–West test for nested models ($p = 0.007$)
  - But not the joint SPA test (superior predictive ability, $p = 0.130$); both are explained below
- TimesFM-2.5 is significantly **worse** than zero
  - S&P 500: R^2 −3.35%, DM $p = 0.001$
  - Forecast s.d. 0.141%, three times the LSTM's
- BET: LSTM R^2 +1.15% (DM p two-sided 0.014, one-sided 0.007)
  - A small signal, tested below for nesting, multiplicity and stale prices

MFM_ch14_returns ▸ Interactive

Nested Models: Clark–West against DM

- **The problem**: zero is nested in the historical mean and AR(1)
  - Under H_0 the larger model only adds estimation noise, so DM rejects too rarely
- Clark–West (Clark & West (2007)): add back the noise term and test its mean, one-sided HAC t-test

$$\tilde{d}_t = e_{\mathbf{0},t}^2 - \left[e_{\mathbf{1},t}^2 - (\hat{r}_{\mathbf{0},t} - \hat{r}_{\mathbf{1},t})^2\right]$$

  - $\hat{r}_{\mathbf{0},t}$: zero forecast (small model); $\hat{r}_{\mathbf{1},t}$: forecast of the large model; $e_{\mathbf{0},t}$, $e_{\mathbf{1},t}$: their errors $r_t - \hat{r}_{\cdot,t}$
  - $(\hat{r}_{\mathbf{0},t} - \hat{r}_{\mathbf{1},t})^2$: the cost of estimating coefficients that are zero under H_0
  - H_0: $E[\tilde{d}_t] = 0$; a large positive t: the large model adds information
- One-sided p-values, DM against Clark–West
  - both tests need a finite, non-zero long-run variance of the differential (for CW against the zero forecast, $\tilde{d}_t = 2r_t\hat{r}_t$)
  - historical mean, S&P 500: DM 0.048, Clark–West 0.007
  - historical mean, BET: DM 0.041, Clark–West 0.003
  - AR(1), S&P 500: DM 0.320, Clark–West 0.055 (the BET AR(1) and rolling windows: Appendix)
- Reading: the correction sharply lowers the p-value of the historical mean in both markets
  - But the gain stays small (R^2 of 0.15% on the S&P 500 and 0.30% on the BET)

Many Tests: Bonferroni and Holm

- 6 models $\times$ 3 markets $=$ 18 tests
  - The smallest one-sided DM p is the BET LSTM, 0.007
- The problem: at least one $p < 0.05$ would appear with probability $1 - 0.95^{18} = 0.60$
  - If no model had a signal and the tests were independent
- Bonferroni: with M tests, reject only if $p < 0.05/M$ (here 0.0028)
- Holm: the same idea, applied step by step from the smallest p
- Both control the probability of any false rejection
- Holm-adjusted p over the 18 tests: 0.130, not significant

Many Tests: Romano–Wolf and SPA

- Within the six BET models, tests that use the correlation between the models
  - Romano–Wolf (Romano & Wolf (2005)): step-down adjusted p of the BET LSTM 0.016
  - SPA test (superior predictive ability, Hansen (2005)), H_0: no model beats zero
  - SPA p-values: BET 0.0185, S&P 500 0.130, Bitcoin 0.354
  - SPA and Romano–Wolf need finite, non-zero long-run variances of the differentials and a valid stationary bootstrap
- "Some model beats zero on the BET" holds if the BET was chosen in advance as the test family
- Holm-corrected across the three markets, the SPA p-value of the BET becomes $3 \times 0.0185 = 0.056$
  - Just above 5%
- "One particular comparison out of 18 is real" does not hold
  - Fix the test family before looking at the results
- Type of evidence: comparison of procedures (it shows which method works better on the same data, not why)
  - On the S&P 500 a single month, March 2020, decides the ranking (Further Topics)
- How Holm, Romano–Wolf and SPA work, with a three-test example, in the Appendix

Test 2: the Question and the Choices

- **Question**: do zero-shot foundation models forecast the daily variance of the S&P 500 better than log-HAR?
  - Log-HAR: the standard econometric benchmark
- Why SPY: the S&P 500 ETF (Exchange-Traded Fund) is very liquid and has 5-minute prices
  - The realised variance is computed from them
- Why log-HAR: it approximates long memory with three averages and only four coefficients (Corsi (2009))
- Why QLIKE: it ranks forecasts correctly (Patton (2011))
  - Even when the true variance is replaced by a noisy proxy
- Horizon: one day, the same as for the VaR of Test 3

Realised Variance, HAR and the QLIKE Loss (1/2)

- **Realised variance** (RV, Andersen et al. (2003)): sum of squared 5-minute log returns within the trading day

$$\mathrm{RV}_t = \sum_{j=\mathbf{1}}^{\mathbf{78}} r_{t,j}^{\mathbf{2}}$$

  - $r_{t,j}$: log return of the j-th 5-minute interval of day t
  - $r_{t,\mathbf{1}} = \ln(C_{t,\mathbf{1}}/O_{t,\mathbf{1}})$: open and close of the 09:30 bar; then $r_{t,j} = \ln(C_{t,j}/C_{t,j-\mathbf{1}})$
  - 78 intervals in the regular SPY session (09:30–16:00, New York time); days without the full grid are dropped
  - Average volatility 12.1% p.a.

- **HAR** (Heterogeneous AutoRegressive, model) (Corsi (2009)): today's RV regressed on yesterday's, the last week's and the last month's averages

$$\mathrm{RV}_t = \beta_{\mathbf{0}} + \beta_d \mathrm{RV}_{t-\mathbf{1}} + \beta_w \overline{\mathrm{RV}}_{t-\mathbf{5}:t-\mathbf{1}} + \beta_m \overline{\mathrm{RV}}_{t-\mathbf{22}:t-\mathbf{1}} + \varepsilon_t$$

  - $\overline{\mathrm{RV}}_{t-\mathbf{5}:t-\mathbf{1}}$, $\overline{\mathrm{RV}}_{t-\mathbf{22}:t-\mathbf{1}}$: averages over the last 5 and 22 trading days
  - β_d, β_w, β_m: daily, weekly and monthly weights

Realised Variance, HAR and the QLIKE Loss (2/2)

- log-HAR: the same regression on $\log \mathrm{RV}$; forecast $\exp(\hat{y} + \hat{s}^2/2)$
  - $\hat{y}$: fitted log RV; $\hat{s}^2$: variance of the log residuals
  - The factor $e^{\hat{s}^2/2}$ turns a log forecast into the mean of RV if the log error is conditionally Normal
- **QLIKE** (quasi-likelihood) loss (Patton (2011)) of a variance forecast $\hat{h}_t$

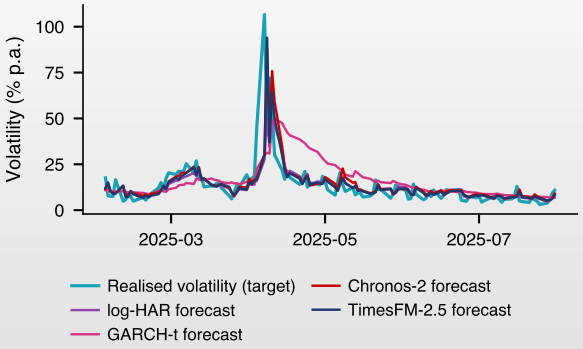
$$L = \frac{\mathrm{RV}_t}{\hat{h}_t} - \log \frac{\mathrm{RV}_t}{\hat{h}_t} - 1$$

  - $L \geq 0$, and $L = 0$ when $\hat{h}_t = \mathrm{RV}_t$
  - Under-forecasts cost more than over-forecasts of the same size
  - Minimised by the conditional mean
  - Ranks forecasts as the true variance σ_t^2 would when RV is a conditionally unbiased proxy, $E[\mathrm{RV}_t \mid \mathcal{F}_{t-1}] = \sigma_t^2$, with $\mathcal{F}_{t-1}$ the information up to $t-1$ (so does MSE)

Design of the Volatility Test

- SPY realised variance, one-day-ahead forecasts from 11.10.2022 to 18.09.2026 (979 days)
  - HAR and log-HAR on a rolling 500-day window
  - Random walk: $\hat{h}_t = \mathrm{RV}_{t-1}$
  - GARCH-t (GARCH(1,1) with Student-t innovations) on daily SPY returns, rescaled to RV units by the ratio of average RV to average squared return over 500 days
  - LSTM on the last 22 values of log RV, retrained every 125 days on 500 days (hidden size 8, three seeds)
- Foundation models on $\log \mathrm{RV}$ (context 512): forecast $\hat{h}_t = \int_0^1 \exp(q_u)\, du$ on the quantile grid
  - q_u are quantiles of log RV, $\exp(q_u)$ those of RV, and their average is the mean of RV
  - On the grid: trapezoid rule over the returned levels, with q_u held at its value at the lowest level below it and at the highest level above it
  - Same rule for all three models, without the TimesFM mean output
  - Post-release window: 218 days from 3 November 2025 (the S&P 500 has 220: the half-day sessions of 28.11 and 24.12.2025 lack the full 5-minute grid)

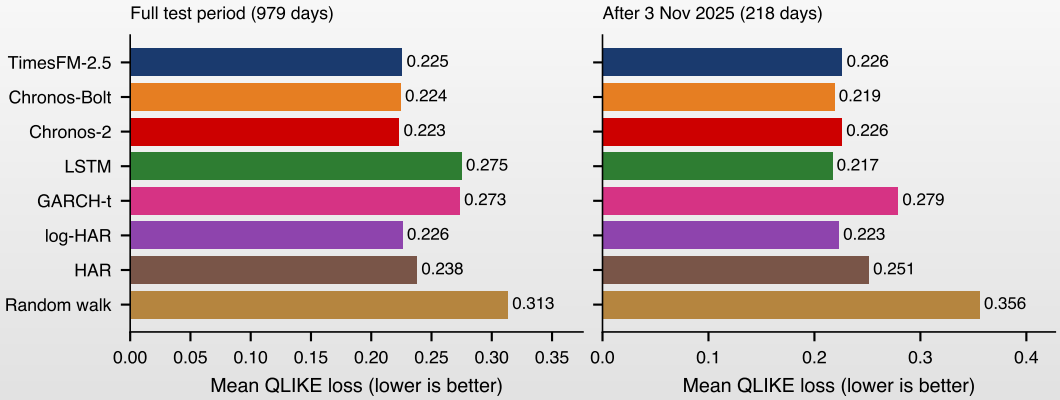
Forecasts Through the April 2025 Shock



- SPY realised volatility (annualised) and one-day-ahead forecasts, February–July 2025
  - Peak 107% p.a. on 07.04.2025, after the United States tariff announcements
- No model anticipates the jump on 04.04.2025 (46% p.a.)
  - Forecasts for 07.04.2025 rise after this jump, to 27–30% p.a.
  - All underestimate that day's 107% peak
  - One-day forecasts extrapolate the past, shocks come from news
- Foundation models and log-HAR decay back at similar speeds
- GARCH-t reacts to daily returns and decays more slowly

MFM_ch14_realised_vol ▶ Interactive

QLIKE Losses of Eight Volatility Models (1/3)



MFM_ch14_realised_vol ▶ Interactive

QLIKE Losses of Eight Volatility Models (2/3)

- Bars are average QLIKE losses
  - Left: full test period (979 days)
  - Right: after 3 November 2025 (218 days)
- Chronos-2 (0.223), Chronos-Bolt, TimesFM-2.5 and log-HAR (0.226): DM tests not significant
  - Chronos-2 $p = 0.539$
  - DM and MCS need finite moments of the differential $d_t = \mathrm{RV}_t(1/h_{A,t} - 1/h_{B,t}) + \ln(h_{A,t}/h_{B,t})$, so also forecasts that do not approach zero
- The 90% model confidence set (MCS, Hansen, Lunde & Nason (2011)) drops the worst model one at a time
  - Until equal average losses are no longer rejected
- The MCS keeps six of the eight models: only the random walk and GARCH-t are eliminated
- HAR in levels over-forecasts (mean forecast / mean RV = 1.35)
- After release the small LSTM is best (0.217): rankings are unstable
- Full table with DM and MCS p-values in the Appendix
- Type of evidence: comparison of procedures

MFM_ch14_realised_vol

QLIKE Losses of Eight Volatility Models (3/3)

- Giacomini–White test (Giacomini & White (2006)): can the loss differential d_t be predicted from yesterday's information?

$$\mathrm{GW} = n\,\bar{Z}^{\top}\hat{\Omega}^{-1}\bar{Z} \;\to\; \chi^2_2, \qquad Z_t = (1,\; d_{t-1})^{\top} d_t$$

  - d_t: QLIKE loss of the model minus that of log-HAR
  - Instruments 1 and d_{t-1}, known at $t-1$
  - H_0: $E[d_t] = 0$ and $E[d_{t-1}d_t] = 0$ jointly
  - $\bar{Z}$: mean of Z_t; n: number of forecasts
  - $\hat{\Omega}$: long-run variance of Z_t (HAC, Bartlett kernel, 6 lags), since $E[Z_t] = 0$ does not rule out autocorrelation in Z_t
- Result: $p = 0.828$ (Chronos-2) and 0.812 (TimesFM-2.5): no conditional difference
  - The test needs a finite, non-singular long-run variance of Z_t
- No difference detected is not proof of equality (power is discussed in the AI section)

Median or Mean under QLIKE

Chronos-2 and TimesFM-2.5 on log RV	QLIKE	Mean forecast / mean RV
Chronos-2, $\int_0^1 \exp(q_u)\,du$	0.223	1.02
Chronos-2, $\exp(q_{0.5})$	0.249	0.80
TimesFM-2.5, $\int_0^1 \exp(q_u)\,du$	0.225	0.94
TimesFM-2.5, $\exp(q_{0.5})$	0.250	0.80

- A theorem (Gneiting (2011)): QLIKE is consistent for the mean
  - The forecast h that minimises its expected value given the information $\mathcal{F}$ is the conditional mean, $\arg\min_h E[L(\mathrm{RV}, h) \mid \mathcal{F}] = E[\mathrm{RV} \mid \mathcal{F}]$
  - Monotone maps preserve quantiles: exp(median of $\log \mathrm{RV}$) is the median of RV, elicited by absolute error (it minimises its mean), not by QLIKE
  - Lognormal RV, s^2 the variance of log RV: median/mean $= e^{-s^2/2}$
  - The median version is significantly worse (DM +3.53 for Chronos-2) and worse than log-HAR (0.226)

Design of the Tail-Risk Test (1/2): Question and Targets

- **Question**: do foundation models give better VaR 1% and ES 2.5% forecasts than FHS?
  - Either directly or as volatility filters
- Targets: VaR_α and ES_α, computed from the quantiles $q_u(r_{t+1} \mid \mathcal{F}_t)$
  - Definitions: slide From a Quantile Grid to VaR and ES
  - $q_u(r_{t+1} \mid \mathcal{F}_t)$: the u-quantile of tomorrow's return given today's information $\mathcal{F}_t$
  - VaR 1%: the level of regulatory Basel backtesting (coverage tests)
  - The pair (VaR 2.5%, ES 2.5%): the ES level of the FRTB standard (Fundamental Review of the Trading Book)
- Classical models: historical simulation (HS), GARCH(1,1)-t, filtered historical simulation (FHS, McNeil & Frey (2000))
  - 1,000-day window, re-estimated every 20 days
- Why FHS as the benchmark
  - It combines a conditional GARCH volatility with the empirical shape of the tails
  - Without assuming a distribution

Design of the Tail-Risk Test (2/2): the Foundation Models

- **Raw**: Chronos-2 quantiles of tomorrow's return (context: 512 returns)
- **Hybrid**: scale $\hat{\sigma}_t$ = forecast of $|r_t|$ (Chronos-2: grid mean; TimesFM: median)
  - The foundation model reads the series $|r|$: context = the 512 absolute returns before day t
  - Then FHS on $r_t/\hat{\sigma}_t$: the empirical quantile of the standardised returns, times $\hat{\sigma}_t$
- Median or mean matters little here, unlike in Test 2: FHS is scale-free

$$c\hat{\sigma}_t \times q_\alpha\big(r/(c\hat{\sigma})\big) = \hat{\sigma}_t \times q_\alpha\big(r/\hat{\sigma}\big)$$

- A constant factor in $\hat{\sigma}$ (such as a stable median/mean ratio) cancels
  - Only a time-varying factor changes the VaR

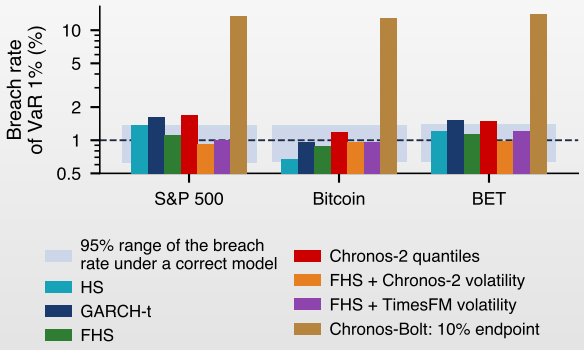
Scoring (VaR, ES) Pairs: the FZ0 Loss

- A functional is elicitable if some loss is minimised, in expectation, by its true value
  - ES alone is not, the pair (VaR, ES) is (Fissler & Ziegel (2016))
- FZ0 loss of a forecast pair, with the daily loss $L_t = -r_t$

$$\mathrm{FZ0} = \frac{1\{L_t > v\}(L_t - v)}{\alpha\, e} + \frac{v}{e} + \log e - 1$$

  - FZ0: the zero-homogeneous Fissler–Ziegel loss, Patton, Ziegel & Chen (2019)
  - v: VaR forecast; $e > 0$: ES forecast; α: tail level; $1\{L_t > v\}$: 1 on a breach day
  - Lower is better: the first term penalises breaches, $\log e$ penalises wide forecasts
- Models compared by DM tests on FZ0 differences and by the MCS of Test 2

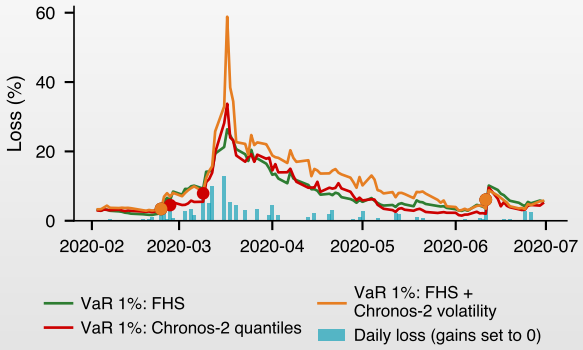
VaR 1% Breach Rates



- Share of days with a loss above VaR 1% (log scale)
  - Light-blue band: 95% range if the model were correct
  - Kupiec (Chapter 8): likelihood-ratio (LR) test that the breach rate equals α, χ^2_1
- Raw Chronos-2: too many breaches on the S&P 500 (1.67%, Kupiec $p = 0.001$) and the BET (1.50%)
- Chronos-Bolt, lowest quantile 10% used as "VaR 1%": 13.5%
- Hybrids: breach rates close to 1% in all three markets
  - Average coverage only
  - Full table in the Appendix

MFM_ch14_tail_risk ▶ Interactive

The COVID-19 Crash: VaR 1% of Three Models



- S&P 500, February–June 2020 (104 days): daily losses and VaR 1% forecasts
  - Dots: breaches
- Breaches: FHS 2, raw Chronos-2 4, FHS with Chronos-2 volatility 2, HS 12
  - All models are hit in the first days
- Chronos-2 reads the context fast, so its quantiles widen quickly
- The hybrid overshoots: VaR 1% of 58.9% on 17.03.2020
  - Against 26.4% for FHS and a largest loss of 12.8% (Appendix)

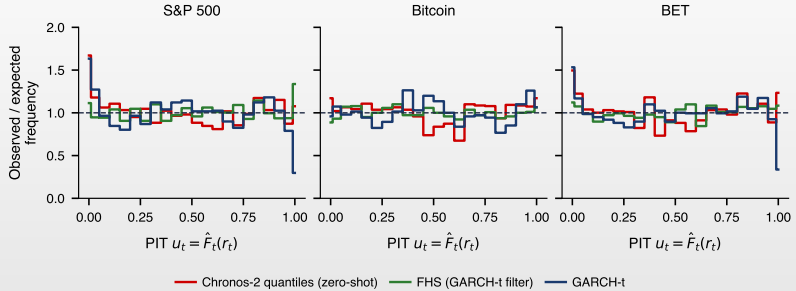
MFM_ch14_tail_risk ▶ Interactive

Backtesting Beyond Kupiec: Clustering and the DQ Test

p: indep. / DQ	HS	GARCH-t	FHS	Chronos-2 raw	FHS + Chronos-2	FHS + TimesFM
S&P 500	0.001 / <0.001	0.205 / 0.001	0.045 / 0.021	0.223 / <0.001	0.021 / 0.001	0.029 / <0.001
Bitcoin	0.611 / 0.760	0.261 / 0.463	0.220 / 0.092	0.404 / 0.034	0.026 / 0.004	0.261 / 0.182
BET	0.399 / <0.001	0.258 / 0.082	0.409 / 0.650	0.270 / 0.026	0.475 / <0.001	0.379 / <0.001

- Independence (Christoffersen (1998)): LR test that a breach today does not change the chance of one tomorrow
  - First-order Markov chain, $\sim \chi^2_1$
  - Conditional coverage = Kupiec + independence, χ^2_2
- Dynamic quantile (DQ, Engle & Manganelli (2004)): $\mathrm{Hit}_t = \beta_0 + \sum_{j=1}^{4} \beta_j \mathrm{Hit}_{t-j} + \beta_5 \mathrm{VaR}_t + u_t$
  - With $\mathrm{Hit}_t = I_t - \alpha$ and $I_t = 1$ on a breach day
  - Wald statistic that all six coefficients are zero, $\sim \chi^2_6$
- S&P 500: the hybrids have the right breach rate but fail independence and DQ
  - Every model fails DQ at 5%
- BET: FHS passes both tests, the hybrids fail DQ
  - Right on average is not right conditionally

The Probability Integral Transform (PIT) and Calibration (1/2)



- PIT (Probability Integral Transform) $u_t = \hat{F}_t(r_t)$: the forecast probability of a return at or below the one realised (Diebold, Gunther & Tay (1998))
  - If the predictive distribution function $\hat{F}_t$ is correct, the u_t are i.i.d. uniform on $(0, 1)$
- Bars: observed over expected frequency in the 22 intervals of the Chronos-2 grid
  - A bar above 1 means too many outcomes in that region

MFM_ch14_calibration ▶ Interactive

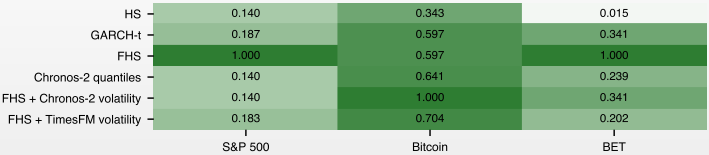
The Probability Integral Transform (PIT) and Calibration (2/2)

- Berkowitz test (Berkowitz (2001)): transform $z_t = \Phi^{-1}(u_t)$, standard Normal if the forecast is correct
  - Test $\mu = 0$, $\sigma = 1$ for z_t, censored outside [1%, 99%]
  - Φ^{-1}: standard Normal quantile function; μ, σ: mean and standard deviation of z_t; $\hat{\sigma} > 1$: tails too thin in the forecast
  - Chronos-2 rejected on the S&P 500 ($p = 0.019$, $\hat{\sigma} = 1.03$) and the BET; FHS never
- Quantile score (QS): the average pinball loss over the K levels u_k of the grid

$$\mathrm{QS} = \frac{2}{K} \sum_{k=1}^{K} \left(1\{r_t < q_{u_k}\} - u_k\right)\left(q_{u_k} - r_t\right)$$

  - A grid analogue of the CRPS (continuous ranked probability score, Gneiting & Raftery (2007))
  - q_{u_k}: forecast quantile at level u_k; lower is better
  - On the S&P 500 and the BET, Chronos-2 is worse than FHS in the centre and in the tails (S&P 500 DM +2.51 and +3.65)
  - On Bitcoin the difference is not significant; table in the Appendix

Comparing (VaR 2.5%, ES 2.5%) Forecasts: the MCS



- p-values of the 90% model confidence set (Hansen, Lunde & Nason (2011)) for the FZ0 loss (darker green: higher)
  - The only model eliminated ($p < 0.10$) is HS on the BET ($p = 0.015$, the near-white cell)
  - FZ0 table in the ▸ Appendix
- S&P 500: FHS has the lowest FZ0
  - The hybrids have the right breach rate but a larger VaR
  - Average VaR 1% over the 2 693 test days: 3.20% for the Chronos-2 hybrid vs 2.78% for FHS
  - Breach clustering does not separate them: 2 back-to-back breaches for FHS and for each hybrid
- Family fixed in advance: the five DM tests against FHS on each market, with the Holm correction
  - S&P 500: FHS beats the Chronos-2 hybrid (DM +2.65, Holm $p = 0.032$)
  - BET: DM +2.36, Holm $p = 0.054$, so superiority is not established at 5%
  - Over the 15 tests of the three markets: Holm $p = 0.096$ and 0.180
- Type of evidence: comparison of procedures

MFM_ch14_tail_risk ▸ Interactive

After the Release: Few Breaches, Little Power

Breaches after 3.11.2025	HS	GARCH-t	FHS	Chronos-2 quantiles	FHS + Chronos-2 vol.	FHS + TimesFM vol.
S&P 500 (220 days)	0	2	2	2	2	2
Bitcoin (320 days)	3	3	4	4	3	4
BET (217 days)	4	4	4	5	1	2

- Expected breaches: 2.2 (S&P 500), 3.2 (Bitcoin), 2.2 (BET)
  - The weights used were public before 3.11.2025 (pinned Hugging Face revision, SHA-256 hash), but the tests have little power
  - Kupiec rule with the asymptotic χ^2_1 cut-off (nominal 5%), power computed exactly from the binomial, against a true rate of 1.67%: 0.19 at $T = 220$ days
  - Actual size (rejection rate when the model is right) 0.13; power 0.50 at $T = 1{,}000$; 0.87 at $T = 2{,}693$
  - 80% power first at about 2 130 days (8.4 years), for good from 2 320
  - Against 2%: power 0.29 at $T = 220$ (Seminar 14, exercises A2 and A9)
- Low power is not zero power
  - HS has 0 breaches on the S&P 500, and $x = 0$ is in the Kupiec rejection region at $T = 220$
  - A non-rejection does not establish calibration

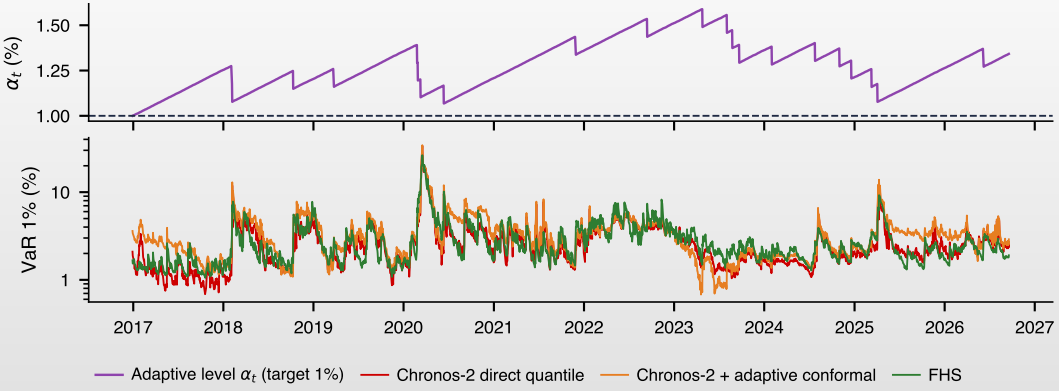
Recalibrating Quantiles: Adaptive Conformal Inference

- Idea: shift the model's quantile by the error it made recently
  - And adapt the level after each breach
- Score $E_t = \hat{q}_{\alpha,t} - r_t$
  - Corrected quantile $\hat{q}_{\alpha,t} - \hat{Q}_{\mathbf{1}-\alpha_t}$
  - $\hat{q}_{\alpha,t}$: the model's α-quantile for day t; $E_t > 0$: the return fell below it
  - $\hat{Q}_{\mathbf{1}-\alpha_t}$: the empirical $(1-\alpha_t)$ quantile of the last 250 scores
- Adaptive level of adaptive conformal inference (ACI, Gibbs & Candès (2021))

$$\alpha_{t+\mathbf{1}} = \alpha_t + \gamma(\alpha - \mathrm{err}_t), \qquad \mathrm{err}_t = 1\{r_t < \hat{q}_{\alpha,t} - \hat{Q}_{\mathbf{1}-\alpha_t}\}, \qquad \alpha_{\mathbf{1}} = \alpha$$

  - α: target level (1%); $\mathrm{err}_t = 1$ on a breach
  - $\gamma > 0$: step size, here 0.001; $\gamma = 0$: rolling calibration at a fixed level (the correction is recomputed daily from the last 250 scores), not standard split conformal
  - After a breach α_t falls (the next VaR is wider); after a calm day it rises slowly

Adaptive Conformal VaR 1% on the S&P 500 (1/2)



MFM_ch14_calibration ▶ Interactive

Adaptive Conformal VaR 1% on the S&P 500 (2/2)

- Top: adaptive level α_t (target 1%), mostly above 1%
  - The score shift alone would over-widen the VaR (with $\gamma = 0$, rolling calibration at a fixed level: 16 breaches, average VaR 1% 3.58%)
  - So ACI loosens the level
- Bottom: VaR 1% of raw Chronos-2, Chronos-2 with ACI, FHS (log scale)
- ACI brings the breaches near 1% (S&P 500: 21 vs 24.4 expected)
  - Conditional coverage is not rejected in any market
- DQ still rejects on the S&P 500 and the BET
  - The VaR is wider (average VaR 1% over the 2 443 days after the 250 calibration days: 3.36% vs 2.85% for FHS on the same days)
  - FZ0 is significantly higher
- The guarantee concerns the average of err_t, not the value of α_t
- Type of evidence: comparison of procedures (ACI against raw Chronos-2 and FHS)
- Proper scoring rules reward sharpness subject to calibration (Gneiting & Raftery (2007))
  - Calibration is necessary, not sufficient

MFM_ch14_calibration

Case Study: Is Complexity a Virtue in Return Prediction? (1/2)

- **Questions** (Kelly, Malamud & Zhou (2024))
  - must a return-prediction model have far fewer parameters P than observations T?
  - can models with complexity $c = P/T > 1$ do better out of sample?
- Setting (eqs. 1–2): next month's return is an unknown function of today's predictors

$$R_{t+\mathbf{1}} = f(G_t) + \varepsilon_{t+\mathbf{1}}, \qquad f(G_t) \approx \sum_{i=\mathbf{1}}^{P} S_{i,t}\,\beta_i$$

  - Approximated by a wide network
  - $R_{t+\mathbf{1}}$: market excess return; G_t: vector of predictors known at t; $\varepsilon_{t+\mathbf{1}}$: unpredictable part
  - $S_{i,t}$: the i-th non-linear feature of G_t (a hidden unit of a one-hidden-layer network); β_i: its weight
  - Universal approximation (Hornik, Stinchcombe & White (1989))
- **Why it matters**: a theoretical case for the large models of this chapter
  - Against the parsimony rule of classical econometrics

Case Study: Is Complexity a Virtue in Return Prediction? (2/2)

- **Link to Test 1**: no model beats the zero forecast on the S&P 500 and Bitcoin
  - After the multiple-testing correction
  - The paper asks whether a negative out-of-sample R^2 means no economic value
  - Economic value: Sharpe ratio of the market-timing return $\pi_t R_{t+1}$, with position $\pi_t = \hat{\beta}' S_t$ (eq. 6)
  - The forecast itself is the position
  - $\hat{\beta}$: estimated weights; S_t: vector of the P features at t
- **Theory** (Sections III–IV): expected out-of-sample R^2 and timing Sharpe ratio rising with c
  - For a misspecified model with an optimally chosen ridge penalty z (shrinks the weights $\hat{\beta}$ towards zero), under the assumptions of the paper
  - Near $c = 1$ (the interpolation boundary) least squares fits the training data exactly and its variance explodes
  - Beyond it, performance recovers

Case Study: Data and Method

- **Data** (Section V.A): monthly excess return of the CRSP value-weighted index, 1926–2020
  - CRSP: Center for Research in Security Prices
  - The 15 predictors of Welch & Goyal (2008) (14 variables and the lagged market return)
  - Returns scaled by their trailing 12-month volatility, predictors by an expanding-window standard deviation
  - 36 months of warm-up, so the sample starts in 1930
- **Random Fourier features** (RFF; eq. 20, Rahimi & Recht (2007))

$$S_{i,t} = \big(\sin(\gamma\,\omega_i' G_t),\ \cos(\gamma\,\omega_i' G_t)\big)', \qquad \omega_i \sim N(0, I), \quad \gamma = 2$$

  - Each feature is a sine or cosine of a random combination of the predictors
  - ω_i: random weight vector, drawn once from the standard Normal distribution (I: identity matrix) and then fixed
  - γ: degree of non-linearity, larger γ gives faster oscillating features
  - A wide two-layer network whose first-layer weights are random; only the output weights β are estimated
  - P counts columns: $P/2$ vectors ω_i, each giving a sine and a cosine (so $P = 12\,000$ uses $6\,000$ vectors ω_i)

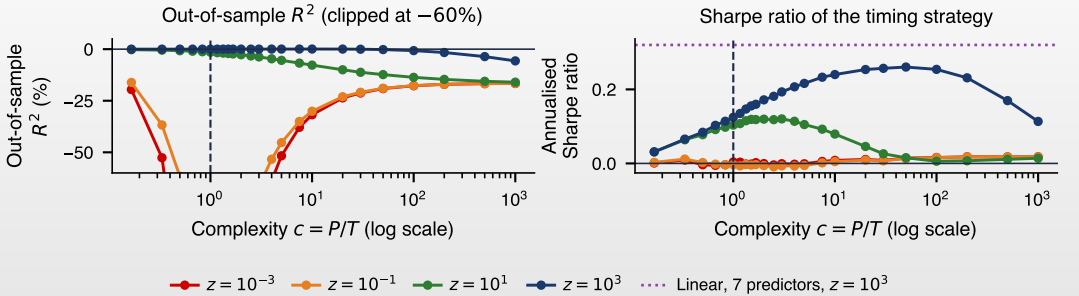
Case Study: Key Results

Model ($T = 12$)	R^2 (%)	Sharpe [t]	Inf. ratio [t]	Max loss	Skewness
Linear, $z = 0^+$	< -100	-0.11 [-1.0]	-0.16 [-1.6]	98.5	-0.9
Linear, $z = 10^3$	-3.8	0.46 [4.4]	0.33 [3.1]	2.4	-0.1
RFF, $c = 1\,000$, $z = 10^3$	0.6	0.47 [4.5]	0.31 [2.9]	1.2	2.5

Source: Table I, Panel A, of the paper; information ratio against the market; maximum monthly loss in standard deviations; the t-statistics test the mean strategy return (Sharpe column) and alpha (information-ratio column)

- Kitchen sink (all predictors), no ridge penalty ($P = 15 > T = 12$ months): R^2 below -100%, Sharpe ratio -0.11
  - Reading: at the interpolation boundary forecasts explode, reproducing the negative verdict of Welch & Goyal (2008)
- Same predictors with $z = 10^3$: $R^2 = -3.8\%$, yet Sharpe ratio 0.46 ($t = 4.4$)
  - Reading: R^2 penalises the scale of forecasts; timing profits depend on their correlation with returns (footnote 5)
- RFF, $c = 1\,000$: information ratio 0.26 ($t = 2.5$) against the best linear model
  - With $T = 60$ and 120, 0.25 and 0.24 against the market ($t = 2.3$ and 2.2; Panels B–C)
  - Reading: nonlinearity extracts more from the same predictors, with a smaller maximum loss and positive skewness
- Positions (Figure 10): almost always long
  - Cut before 14 of 15 NBER (National Bureau of Economic Research) recessions, out of sample
  - Reading: the model learns the sign restriction of Campbell & Thompson (2008) without being given it
- Type of evidence: comparison of procedures (models of different complexity, out of sample)

Case Study: Complexity on Course Data (1/2)



MFM_ch14_virtue_complexity ▶ Interactive

Case Study: Complexity on Course Data (2/2)

- Figures 7–8 of Kelly, Malamud & Zhou (2024) on course data: SPY excess returns, 7 predictors
  - 354 test months (February 1997–August 2026)
  - $T = 12$ complete observations, $\gamma = 2$, 1 000 draws
  - Dashed: $c = 1$; dotted: linear, $z = 10^3$
  - Data in the Appendix
- Nearly ridgeless ($z = 10^{-3}$): $R^2 = -1\,852\%$ at $c = 1$, recovering to -16.6% at $c = 1\,000$
  - The double descent of Figure 7 (left panel cut at -60%)
- $z = 10^3$: Sharpe ratio 0.12 at $c = 1$, 0.26 at $c = 50$, 0.11 at $c = 1\,000$ ($t = 0.6$)
  - Linear with $z = 10^3$: Sharpe 0.32 ($t = 1.7$), not beaten by RFF
  - t tests the mean strategy return ($\sqrt{n}\,\bar{R}/s_R$), not the Sharpe ratio itself
  - With HAC errors (Newey–West, 3, 6 and 12 lags): $t = 1.75$, 1.81 and 1.81
  - Information ratio of RFF 0.07 against the linear model, $t = 0.4$; -0.12 against SPY
- Reading: on course data, the RFF Sharpe ratio rises only up to $c \approx 50$ and never beats the linear model
- Type of evidence: comparison of procedures

MFM_ch14_virtue_complexity

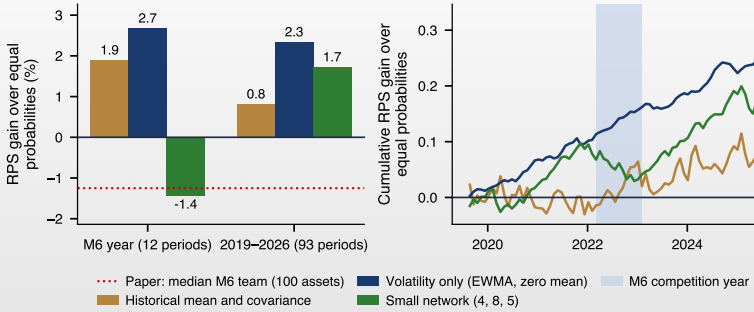
Case Study: Limits and an Open Question

- **Limits** stated by the authors
  - Theory for linear ridge models on fixed random features and a single asset
  - A panel with cross-asset covariances is left open (Introduction, footnote 2)
  - Predictability spread evenly across features (Assumption 4)
  - Sparse LASSO-type (Least Absolute Shrinkage and Selection Operator) models left to future research (footnote 22)
  - No optimal degree of nonlinearity γ
  - Average timing returns about half as large in 1975–2020 as in 1930–1974 (Section V.F)
- **Course data** (Seminar 14, exercise C3): 7 of the 15 predictors and about 30 years
  - The power of a two-sided 5% test for an information ratio of 0.31 is only about 39% ($t \approx 0.31\sqrt{29.5} \approx 1.7$)
- **Open question**: learned instead of random features
  - The embeddings of a pretrained foundation model such as Chronos-2 as S_t
  - Is complexity still a virtue, on a panel of course markets?

Case Study: The M6 Forecasting Competition

- **Question** (Makridakis et al. (2025)): can a forecasting method, however complex, beat equal probabilities in forecasting the relative performance of assets?
  - M6: a live competition, 2022–2023, 226 teams, 100 assets (50 S&P 500 stocks and 50 international ETFs)
  - The task: the probability that the four-week return of each asset falls in each of the five quintiles of the universe
- **Why it matters**: an out-of-sample evaluation in real time, hence free of contamination through pretraining
  - Teams used machine learning, time-series models and their own judgment
- **Main claim**: equal probabilities are hard to beat
  - Of the 163 teams of the global leaderboard, 38 (23.3%) have an RPS (ranked probability score) below the benchmark's 0.16; only 3 beat it in all 12 months
  - The 11 teams that beat the benchmark in both forecasting and investing reduce the RPS by at most 2.2%
  - Among the top 20% of teams by RPS, 14 use machine learning and 12 time-series models (Table 7)
  - The correlation between the RPS and the information ratio (IR) of the team's portfolio is 0.04 ($p = 0.601$), over the 138 teams whose forecasts differ from the benchmark

Case Study: M6 on Course Data



- Left: RPS reduction over equal probabilities (%), in the 12 M6 periods and in 93 four-week periods (26.07.2019–11.09.2026); dotted line: the median M6 team (RPS 0.162, computed in the paper over 162 of the 163 teams; 100 assets), a context marker, not a comparable threshold
- Right: the cumulative RPS reduction, period by period; the blue area marks the competition year

MFM_ch14_m6_competition ▶ Interactive

Case Study: The RPS, the Data and the Replication

- **RPS** (eq. 1): the distance between the forecast and the realised cumulative distribution over the quintiles

$$\mathrm{RPS}_i = \frac{1}{5}\sum_{j=1}^{5}\Big(\sum_{k\le j} q_{ik} - \sum_{k\le j} f_{ik}\Big)^2$$

  - f_{ik}: forecast probability that asset i falls in quintile k; q_{ik} (M6 notation, not a quantile here): in M6, 1 for the realised quintile and 0 otherwise
  - Here, with 19 assets: the share of quintile k covered by the asset's rank interval $[(r-1)/19,\ r/19]$ (fractional weights)
  - With 100 assets, equal probabilities ($f_{ik} = 0.2$) give an RPS of 0.16; with the fractional weights here, 0.1516
- **Adaptation**: the 19 M6 assets in the course data (AMZN, JPM, META, HYG, IEF, IWM, LQD, SHY, TLT and the 10 XL* sector ETFs), total returns from adjusted prices
  - The four-week M6 calendar (Table 1), extended from 2019; with 19 assets the realised quintiles are fractional
  - Historical mean: Normal returns with the mean and covariance of the last 250 days, 20,000 simulations
  - Volatility only: zero mean and the EWMA covariance (exponentially weighted moving average, $\lambda = 0.94$)
  - The network: 4 inputs (volatilities and returns over 20 and 250 days), 8 neurons, 5 probabilities; RPS as the loss, re-estimated each year on the periods since 2004

Case Study: The M6 Result on Course Data

- The volatility-only model beats the benchmark: an RPS lower by 2.7% in the M6 year (11 of 12 periods) and by 2.3% since 2019 (DM -6.35, $p < 0.001$)
  - 2.7% here and 2.2% in the paper come from different universes and targets, so they are not directly comparable
  - The mechanism: in the first M6 period, META (EWMA volatility 76% a year) has a probability of 0.82 of landing in an extreme quintile, HYG (7%) 0.08, the benchmark 0.40
- The historical mean lowers the RPS by 1.9% in the M6 year but only by 0.8% since 2019 ($p = 0.358$)
  - With the same 250-day covariance, a zero mean has an RPS 1.7% lower than the estimated mean (DM -1.75, $p = 0.080$)
- The network raises the RPS by 1.4% in the M6 year and lowers the RPS by 1.7% since 2019 ($p = 0.016$)
  - Against the volatility-only model: an RPS 0.6% higher, a difference not detected (DM 0.73, $p = 0.466$)
  - Without the return inputs the RPS rises by only 0.05% ($p = 0.70$)
- Type of evidence: comparison of procedures; the useful information is in the dispersion of returns, not in their sign, as in Tests 1 and 2
- Limits: 19 assets, not 100, and one market
  - The RPS is bounded, so the DM test needs no assumption on the tails
  - In the paper: a single year, and only 26 of the 163 teams submitted new forecasts at all 12 deadlines
  - The periods before 2022 use, with hindsight, the universe chosen for M6

MFM_ch14_m6_competition

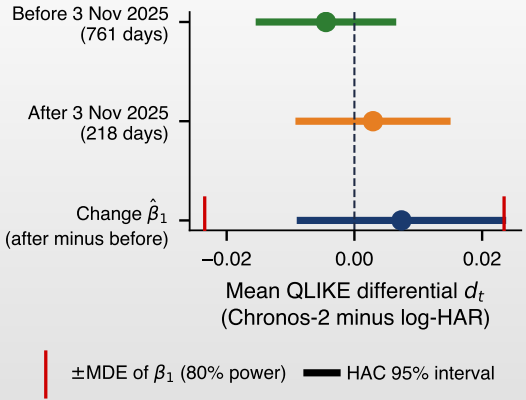
Open Question: Forecasting Skill or Memorisation of Training Data?

- Test 2: zero-shot Chronos-2 is not distinguishable from log-HAR on SPY realised variance
- Rahimikia, Ni & Wang (2025): on daily excess returns across global markets, off-the-shelf foundation models do poorly
  - Zero-shot and after fine-tuning
  - Models pretrained from scratch on financial data gain
- Lopez-Lira, Tang & Zhu (2025): once a model has seen the realised values, its counterfactual forecasting skill is not identified
  - Detected memorisation is only a lower bound
  - Shown for LLMs; the logic applies to any pretrained forecaster
- Open questions
  - is the parity learned structure (long memory, mean reversion) or recall of the test period?
  - can a short post-release window tell?
- A test: $d_t = L(\hat{h}_t^{FM}, \mathrm{RV}_t) - L(\hat{h}_t^{\log \mathrm{HAR}}, \mathrm{RV}_t)$
  - The QLIKE loss L of the foundation model (FM) minus that of log-HAR
  - Memorisation implies $E[d_t \mid \text{post}] > E[d_t \mid \text{pre}]$: the FM loses its edge after the release of its weights
  - H_0: equality; a rejection may also reflect a regime change

The Discovery Loop in Five Steps

- **Literature**: ask an assistant for evaluations of time-series foundation models on financial data since 2024
  - Check every arXiv identifier and read the corpus documentation yourself
- **Hypotheses**: predictions that separate memorisation from skill
  - Pre vs post release, never-public series, simulated series with the same dynamics
- **Data and code**: from the chapter's volatility forecasts, the QLIKE differential Chronos-2 minus log-HAR
  - A pre/post test with HAC errors and the minimum detectable effect (MDE)
- **Robustness**, fixed in advance: a placebo release date one year earlier must show no break
  - Simulated series cannot be memorised
- **Critique**: ask the assistant to attack "no memorisation, because the post-release test is insignificant"
  - Expect low power and different regimes
  - Reusable prompts in the ▸ Appendix

Mini-Case: a Test Without Power (1/3)



MFM_ch14_ai_discovery ▶ Interactive

Mini-Case: a Test Without Power (2/3)

- Typical assistant proposal: "the parity of Chronos-2 comes from memorised data, so its edge must vanish after release"
- Chart: mean QLIKE differential d_t (Chronos-2 minus log-HAR) before and after the release
  - Green: before; orange: after
  - With 95% HAC intervals
- Bottom row: the estimated change $\hat{\beta}_1$, with red lines at $\pm$MDE
  - MDE: the smallest effect the test can detect
- Mean d_t: -0.0045 before 3 Nov 2025 (761 days), $+0.0029$ after (218 days)
- Change $+0.0074$, HAC $t = 0.88$, $p = 0.38$
  - The test needs a finite long-run variance of d_t, so forecasts that do not approach zero
- Type of evidence: association (the variables move together; this does not prove causation)
  - Even a rejection would tie the difference to the release date only, not to memorisation

MFM_ch14_ai_discovery

Mini-Case: a Test Without Power (3/3)

- Minimum detectable effect: MDE $= (z_{0.975} + z_{0.8})\,\mathrm{SE}_{HAC}(\hat{\beta}_1) = 0.0234$
  - $z_{0.975} = 1.96$, $z_{0.8} = 0.84$: standard Normal quantiles, 5% two-sided test, 80% power
  - $\hat{\beta}_1$: estimated post-release shift in the mean of d_t; SE_{HAC}: its HAC standard error
  - 5.3 times the pre-release gap
  - Still 0.0157 with an infinite post-release window
- Why a floor: HAC errors give each period its own long-run variance V

$$\mathrm{SE}^2 \approx \frac{V_{pre}}{761} + \frac{V_{post}}{218}$$

  - More post-release days remove only the second term, leaving $2.80\sqrt{V_{pre}/761}$
- Reading: "no memorisation" is not a finding
  - Power must come from the cross-section (many series tested jointly), not from waiting

MFM_ch14_ai_discovery

Your Turn: a Project Seed

- Questions (where memorisation is impossible)
  - does zero-shot skill for daily variance survive after release?
  - does it survive on simulated series?
  - does it survive on less-covered markets?
- Replicate first: the chapter's QLIKE of log-HAR (0.226) and Chronos-2 (0.223)
  - DM $p = 0.539$
  - The 218-day post-release window
- Deliverables
  - A pre-registered power calculation
  - A panel of series from the course data with a range-based variance proxy and a proxy-robust loss (Patton (2011))
  - Pre/post tests with placebo release dates; the MCS across models
  - A referee-style critique and your reply
- Log each prompt that shaped a decision in `AI_USE.md`
  - Each invented reference, leak or wrong number caught in `AI_ERRORS.md`, with the test that caught it
  - Project templates

Key Takeaways

- LSTM gates can mitigate vanishing gradients
  - With a forget-gate bias of 0 they still vanish
- Attention connects distant time steps directly
- Foundation models forecast zero-shot from a context window and return quantiles
- The three tests
  - Returns: after the multiple-testing correction, nothing beats zero on the S&P 500 and Bitcoin
  - Volatility: foundation models not distinguishable from log-HAR
  - Tails: FHS remains the benchmark; foundation models at most as volatility filters
- Evaluate every new model against a strong benchmark and after the release of its weights
  - With inference that matches the design: Clark–West for nested models
  - SPA or Romano–Wolf for many models
  - PIT, quantile scores and DQ for distributions
- Further reading in the Appendix

Check Yourself, and Next: Chapter 15

Retrieval questions

- Why does a stable RNN forget at the rate of the VAR(1) impulse response A^k, not necessarily like $\rho(A)^k$?
- Why is $\exp(q_{0.5})$ of log RV a biased forecast under QLIKE?
- Why can a hybrid VaR pass Kupiec and still lose the FZ0 comparison?
- Why is "not significant after release" not evidence against memorisation?

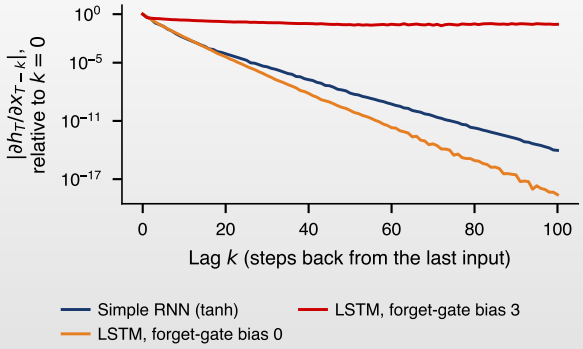
Next: LLMs and sentiment analysis

- The next lecture: Chapter 15
- Can text read by a large language model forecast returns or volatility?
- The same threats return: look-ahead, memorisation of the test period, weak benchmarks
- The same tools: post-release windows, Clark–West, multiplicity control

The Practical Decision

- **What I choose**
  - Daily index returns: the zero forecast
  - Realised volatility: log-HAR; a zero-shot foundation model when thousands of series must be forecast without estimation
  - VaR 1% and ES 2.5%: FHS, with a foundation model at most as a volatility filter or with conformal recalibration
- **Under which conditions**
  - On the S&P 500 and Bitcoin, no model beats the zero forecast after the multiple-testing correction
  - On the BET, the SPA p-value, Holm-corrected for the three markets, is 0.056
  - On SPY, Chronos-2 and log-HAR have QLIKE 0.223 and 0.226 (DM $p = 0.539$)
  - On the S&P 500, FHS has the lowest FZ0 loss (DM +2.65 against the Chronos-2 hybrid)
- **What result would make me drop it**
  - A model that beats the zero forecast after Holm or SPA, in a test family fixed in advance
  - A foundation model with a QLIKE significantly below log-HAR in the post-release window
  - A hybrid VaR that also passes the independence and DQ tests, not only Kupiec

The Vanishing Gradient, Measured



- $|\partial h_T/\partial x_{T-k}|$ by automatic differentiation, relative to $k = 0$ as on the vertical axis
  - 16 hidden units, 50 random initialisations, same Normal inputs
- Simple RNN: 8.3×10^{-5} at lag 20, 7.4×10^{-9} at lag 50
  - Nothing is learned from inputs more than a few weeks back
- LSTM, forget-gate bias 0: 1.2×10^{-10} at lag 50
  - No better than the RNN: $f \approx 0.5$ halves the memory at every step
- LSTM, bias 3: 8.4% at lag 50
  - As predicted by $0.953^{50} \approx 0.09$
  - Initialise the forget-gate bias at a positive value

MFM_ch14_vanishing_gradient

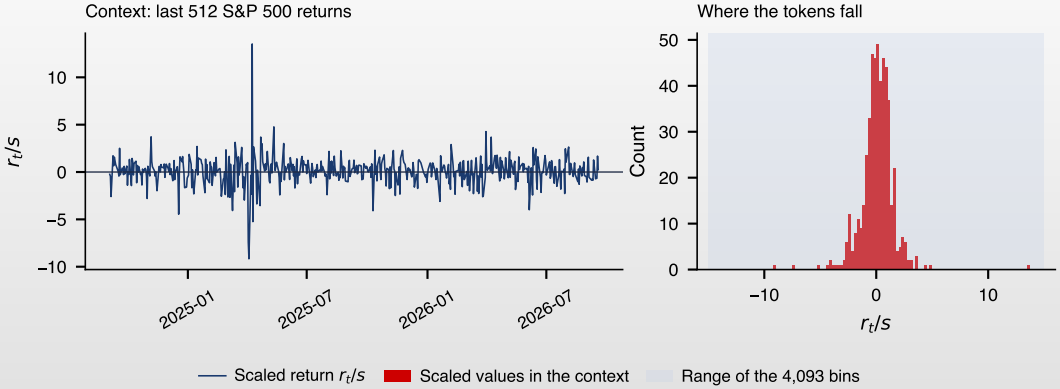
Attention in Numbers: Three Steps and a Causal Mask

- Three time steps, $d_k = 1$: queries $q = (1, 0, 2)$, keys $k = (0.5, 1, -1)$, values $v = (1, 2, 3)$
- Step 3: scores $q_3 k_j = (1, 2, -2)$
  - Softmax weights $e^{s_j} / \sum_i e^{s_i} = (2.718, 7.389, 0.135)/10.243 = (0.265, 0.721, 0.013)$
  - Output $0.265 \cdot 1 + 0.721 \cdot 2 + 0.013 \cdot 3 = 1.75$: step 3 draws mostly on step 2, whose key matches its query best
- **Causal mask**: in a forecaster that predicts every next step, step t may use only steps $j \leq t$

$$\text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right) V$$

  - Equivalently add $M_{tj} = 0$ for $j \leq t$, $M_{tj} = -\infty$ for $j > t$
  - Step 1 then puts weight 1 on itself (output 1)
  - Without the mask its scores $(0.5, 1, -1)$ give weights $(0.35, 0.57, 0.08)$: 57% from step 2, a later day
  - With the mask, step 2 ($q_2 = 0$, equal scores on steps 1 and 2) outputs $(1 + 2)/2 = 1.5$
- Every value in a 512-day context is already observed at the forecast origin
  - So a model reading only the context may attend in both directions
  - The mask matters where the model is trained to predict later positions from earlier ones

Input of the Chronos Model: Scaled Returns and Tokens (1/2)



MFM_ch14_tokenization

Input of the Chronos Model: Scaled Returns and Tokens (2/2)

- Context: the last 512 S&P 500 daily returns
  - Scale $s = 0.672\%$
  - Scaled returns from -9.2 to 13.5, 96% within ± 3
- The 512 returns use 336 distinct tokens out of 4,093
  - Scaling makes the model free of units, so the volatility level must be read from the context
- A crash day with return M (in %) enters the context and the oldest return x_{out} leaves it
  - s rises by $(|M| - |x_{\text{out}}|)/C$ percentage points, approximately $|M|/C$
  - The relative rise, about $|M|/(Cs)$, is large in a short context and small in a 512-day one
  - Seminar 14, exercises A3 and A6

MFM_ch14_tokenization

Forecasting Transformers: the Transformer Block

- **Transformer block** (Vaswani et al. (2017)): multi-head attention, then a small MLP applied to each position separately

$$Z = \mathrm{LN}(X + \mathrm{MHA}(X)), \qquad H = \mathrm{LN}(Z + \mathrm{MLP}(Z))$$

  - Each part has a skip connection and a normalisation
  - $\mathrm{MHA}(X) = [\mathrm{head}_1, \ldots, \mathrm{head}_m]\, W_O$ (multi-head attention): m attentions in parallel, each with its own W_Q, W_K, W_V, concatenated and projected by W_O
  - LN (layer normalisation): each row is standardised, then rescaled by learned parameters
  - $X + \mathrm{MHA}(X)$: the skip connection, which lets the gradient pass straight through the block
  - A model stacks several blocks: H is the input of the next block

Forecasting Transformers: Architectures and Losses

- Three ways to assemble the blocks
  - **Encoder**: attention in both directions inside the context; Chronos-2 is encoder-only and returns the quantile grid in one pass (Ansari et al. (2025))
  - **Decoder-only**: causal mask, trained to predict the next patch at every position: TimesFM (Das et al. (2024))
  - **Encoder–decoder** (T5, used by the original Chronos): the encoder reads the context; the decoder writes the forecast one token at a time
- **Cross-entropy** (original Chronos): predicting the next token is a classification over 4,093 bins

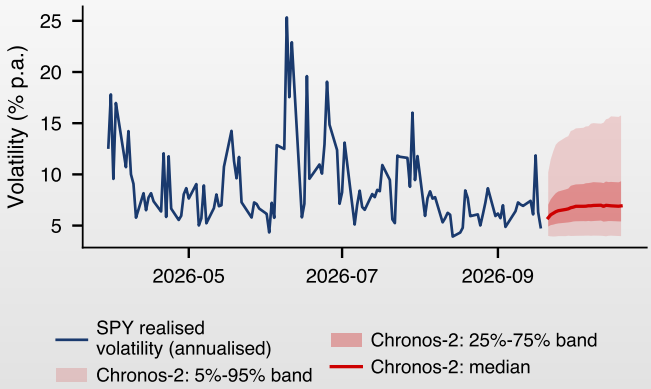
$$\mathcal{L}(\theta) = -\sum_{h=\mathbf{1}}^{H} \log \hat{p}_\theta \left(z_{C+h} \mid z_{\mathbf{1}}, \ldots, z_{C+h-\mathbf{1}} \right)$$

  - $\hat{p}_\theta(z \mid \cdot)$: the probability the model gives to the token that occurred, given the earlier tokens
- **Robust scaling** (Chronos-2, Ansari et al. (2025)): $\tilde{x}_t = \sinh^{-\mathbf{1}} \left((x_t - \hat{\mu})/\hat{\sigma} \right)$
  - $\hat{\mu}$, $\hat{\sigma}$: the mean and standard deviation of the context
  - $\sinh^{-\mathbf{1}}$ grows like a logarithm for large values and so damps outliers

From a Quantile Grid to VaR and ES

- Worked example, S&P 500, quantiles for 18.09.2026 (context up to 17.09.2026): $q_{0.01} = -2.704\%$, $q_{0.05} = -1.400\%$
  - VaR 1% $= 2.704\%$
  - linear interpolation between the two levels: $q_{0.025} = q_{0.01} + \frac{0.015}{0.04}(q_{0.05} - q_{0.01}) = -2.215\%$, so VaR 2.5% $= 2.215\%$
  - ES 2.5%: q_u flat below 1%, trapezoid rule between 1% and 2.5%
- $\mathrm{ES}_{2.5\%} = -\frac{1}{0.025}\left[0.01\, q_{0.01} + 0.015\, \frac{q_{0.01}+q_{0.025}}{2}\right] = 2.56\%$
  - In numbers: $40 \times [0.01 \times 2.704 + 0.015 \times 2.459] = 40 \times 0.06392 = 2.557$, i.e. 2.56%
- Reading: ES 2.5% is close to VaR 1%
  - ES/VaR 2.5% $= 1.15$ is below the Normal value 1.19, although this grid has a heavier left tail than the Normal
  - The flat tail below 1% understates ES (Seminar 14, exercise A7)

A Zero-Shot Fan Chart for Volatility (1/2)



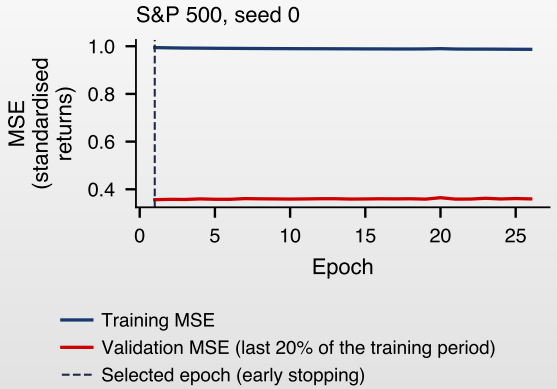
MFM_ch14_fan_chart

A Zero-Shot Fan Chart for Volatility (2/2)

- Chronos-2 on log realised variance of SPY (512-day context)
  - SPY: the exchange-traded fund (ETF) that tracks the S&P 500
  - 22 trading days ahead from 18.09.2026
  - Bands: quantiles 5–95% and 25–75%
- Last volatility 4.8% p.a., 20-day average 6.8%
  - Median forecast at 22 days 6.9%, 90% band 4.0–15.8%
- Mean reversion, as in GARCH and HAR (Chapters 5, 9), with no estimation on SPY
  - The band is skewed upwards partly by construction: volatility is $\sqrt{252\, e^{q}}$ of a log-variance quantile q
  - A convex map stretches the upper side
  - On the log-volatility scale (half the log-variance), the band still lies 0.56 below and 0.82 above the median, so the model itself also leans towards spikes

MFM_ch14_fan_chart

The LSTM for Returns: Training Curves



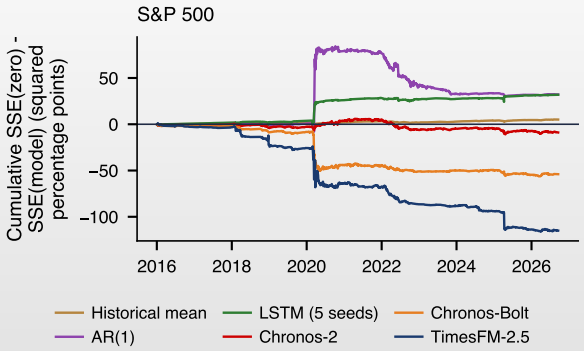
MFM_ch14_returns

The LSTM for Returns: Architecture and Early Stopping

- Input: the last 20 returns, standardised on the training period
  - $h = 16$ hidden units, $d = 1$ input per step, linear output
  - $4(hd + h^2 + 2h) + h + 1 = 1\,233$ parameters (4: the three gates and the candidate; $2h$: two bias vectors per gate)
- MSE, Adam, weight decay 10^{-4}
  - Trained once on up to 4,000 days, the last 20% being the validation block
  - Early stopping: training stops after 25 epochs without a lower validation MSE (patience), and the best weights are kept
  - Forecast = average of 5 seeds
- Chart (S&P 500, seed 0): training and validation MSE by epoch, on standardised returns
  - The dotted vertical line marks the epoch kept
- Training stops 26 epochs in, and the weights of epoch 1 are kept
  - The kept network has seen the training data once and stays close to its random start: one reason why the seed matters
  - Validation MSE is lower than training MSE mainly because the last years of the training period were calmer
- Little to learn: forecast s.d. 0.047% against 1.13% for returns

MFM_ch14_returns

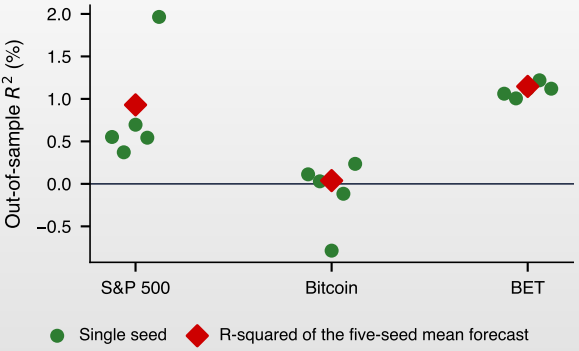
Cumulative Squared-Error Gains on the S&P 500



- Curve SSE(zero) − SSE(model) $= \sum_{s \le t} \left[r_s^2 - (r_s - \hat{r}_s)^2 \right]$
  - SSE: sum of squared errors, in %^{2}
  - Rises when the model beats the zero forecast, falls when it loses
- All curves stay near zero in calm years and jump in March 2020
  - One month decides the ranking
- AR(1) gains in 2020 from the negative autocorrelation of crash days
  - Then gives part of the gain back
- TimesFM-2.5 loses steadily
- Report the path, not only the final R^2

MFM_ch14_returns

The Seed Lottery



- Out-of-sample R^2 of five LSTMs that differ only in the random seed, and of their average
- Single seeds: variation as large as the effect
  - Bitcoin: from -0.79% to $+0.24\%$
  - S&P 500: from $+0.37\%$ to $+1.96\%$
- BET: robust across seeds
  - All 5 seeds between $+1.01\%$ and $+1.22\%$
- Choosing the best seed after seeing the test is backtest overfitting (Chapter 13)

MFM_ch14_returns

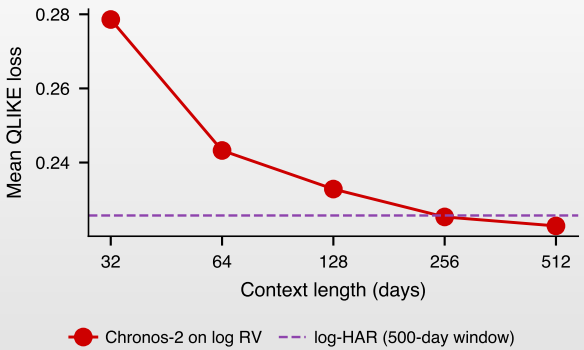
Is BET Predictability Real? Non-Synchronous Trading

- Stale prices of thinly traded constituents make index returns positively autocorrelated (Lo & MacKinlay (1990))
  - Even if every price is efficient
  - First-order autocorrelation since 2016: BET 0.06, S&P 500 −0.14
  - The LSTM forecast has correlation 0.88 with the previous day's BET return: it mostly learned this autocorrelation
- Tradable check: the BET-tracking ETF Patria-TVBETETF, 2 629 common days, same LSTM signal
  - Same day: R^2 2.88% (BET 1.26% on the same 2 629 days)
  - But the ETF is thinly traded too (autocorrelation 0.07, 4.2% zero-return days)
  - Executed one day later: R^2 0.57%, Clark–West $p = 0.003$
- Reading: the evidence is consistent with stale-price effects
  - A small signal survives the execution lag, before transaction costs (Chapter 2)

Computing the Grid Mean: Sensitivity to the Tail Rule

- Toy grid: levels 25%, 50%, 75%, log-quantiles $q = -1, 0, 1$; target $E[Y] = \int_0^1 e^{q_u}\, du$
  - Equal weights on the three points: $(e^{-1} + 1 + e)/3 = 1.36$
  - Course rule (trapezoid, flat below 25% and above 75%): $\frac{1}{4}\left[e^{-1} + \frac{e^{-1}+1}{2} + \frac{1+e}{2} + e\right] = 1.41$
  - A Normal log Y with these quartiles ($s = 1/z_{0.75} = 1/0.674 = 1.48$, $z_{0.75}$ the 75% standard Normal quantile) has $E[Y] = e^{s^2/2} = 3.00$
  - The flat tails miss more than half of it
- Reading: exp is convex and unbounded, so the top of the distribution carries much of the mean
  - Truncating the upper tail can understate the mean badly; the sign of the total error is not guaranteed (a flat lower tail raises the integral)
  - Test 2: mean forecast / mean RV is 1.02 for Chronos-2 (grid to 99%), 0.97 for Chronos-Bolt and 0.94 for TimesFM-2.5 (grids to 90%)
  - Consistent with a cut upper tail, though the forecasts themselves also differ
- Report the rule with the forecasts

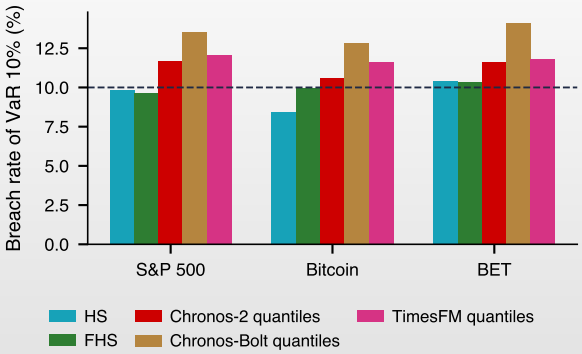
How Much Context Does Chronos-2 Need?



- QLIKE of Chronos-2 on log RV with contexts of 32 to 512 days, on the same forecast days
- QLIKE falls from 0.279 (32 days) to 0.223 (512 days)
  - Below about 256 days Chronos-2 is worse than log-HAR (0.226)
- Long memory: months of history reveal the slow-moving level
  - HAR encodes it explicitly through its 22-day average
- Context length is a modelling choice to report and test, like a window length

MFM_ch14_realised_vol

Where Deciles Suffice: VaR 10%



- Breach rate of the 10% quantile (target 10%) for HS, FHS and the three foundation models
- S&P 500: HS 9.8%, FHS 9.6%
  - Chronos-2 11.7%, Chronos-Bolt 13.5%, TimesFM-2.5 12.0%
- Kupiec rejects the foundation models in eight of nine cases
  - The lower tail sits too high, through location, scale or tail shape
- Plausible reason: pretraining corpora are dominated by smoother series
  - As volatility filters in FHS the models inherit the tails from the data

MFM_ch14_tail_risk

Adaptive Conformal Inference: the Coverage Guarantee and ES

- Guarantee for any sequence of returns, no model needed: over T days the breach rate approaches α

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t - \alpha \right| \leq \frac{\max(\alpha_1, 1-\alpha_1) + \gamma}{\gamma T}$$

  - the bound shrinks like $1/T$: a long-run statement only, loose for small γ
- ES 2.5%: Chronos-2 ES shifted by the VaR 2.5% correction
  - A heuristic without a conformal guarantee
  - Rules and the full table in the Appendix

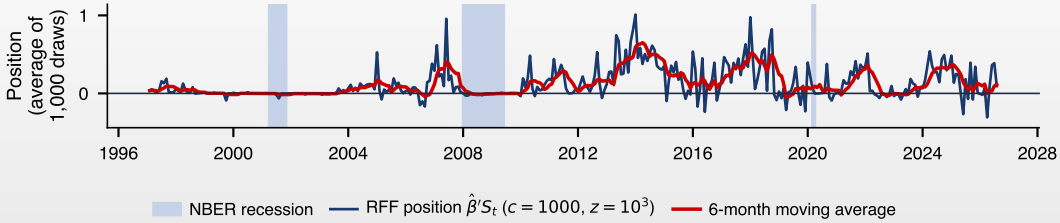
Case Study: the Ridge Estimator, Design and Evaluation

- **Ridge** (Section II.A), no intercept

$$\hat{\beta}(z) = \left(zI + \frac{1}{T} \sum_t S_t S_t' \right)^{-1} \frac{1}{T} \sum_t S_t R_{t+1}$$

  - $z > 0$: ridge penalty, shrinks $\hat{\beta}$ towards zero; $z \to 0^+$: the ridgeless (pseudo-inverse) solution; T: window length
- **Design** (Section V.C): rolling windows of $T = 12$, 60, 120 months
  - P from 2 to 12 000; $\log_{10} z = -3, \ldots, 3$
  - Results averaged over 1 000 independent draws of the features
- **Evaluation**: out-of-sample $R^2 = 1 - \mathrm{Var}(\text{error})/\mathrm{Var}(R)$ (footnote 40)
  - The theory divides the MSE by $E[R^2]$, the zero forecast of Test 1
  - Annualised Sharpe ratio
  - alpha and information ratio (alpha over residual volatility) from a regression of the timing return on the market return
  - Benchmark: the linear "kitchen sink" regression on all 15 predictors (Section V.D)

Case Study: Timing Positions and Recessions (1/2)



MFM_ch14_virtue_complexity

Case Study: Timing Positions and Recessions (2/2)

- On course data, as in Kelly, Malamud & Zhou (2024): position $\hat{\beta}' S_t$ of RFF with $c = 1\,000$, $z = 10^3$
  - Shaded: NBER recessions (series `USREC` from FRED, Federal Reserve Economic Data)
  - Red: 6-month moving average
- Long in 73% of months, average position 0.13
  - A positive equity premium learned without a sign restriction
- Average position 0.05 in the 6 months before the 3 recessions and -0.01 during them
  - Against 0.14 otherwise
  - The 2020 pandemic shock is not anticipated
- Three recessions are a description, not a test: the paper has 15 recessions and 90 years

MFM_ch14_virtue_complexity

Methodological Checks for the Researcher

- **Leakage**: pretraining overlap, full-sample scaling, HAR averages that include day t
  - Seminar 14, exercise B10: they cut QLIKE by a quarter
  - Test: post-release window, simulated or private series
  - Recompute the forecast for t from data truncated at $t-1$
- **Weak benchmarks**: compare with zero (returns), log-HAR (volatility), FHS (tails), not with a random walk
- **Wrong functional**: medians used as means; deciles used as 1% quantiles
  - Mean forecast / mean RV should be near 1; Chronos-2 median: 0.80
- **Selection**: seeds, contexts and models chosen after seeing the test
  - Log every trial, compare the whole set with the MCS
- **Wrong inference**
  - DM on nested models
  - No multiplicity control
  - Coverage without independence
  - "Not significant" reported without the MDE

Norms, Spectral Radius and Explosion

- $\gamma_\varphi \|W\|_2 < 1$ is sufficient for geometric vanishing, so $\gamma_\varphi \|W\|_2 \geq 1$ is necessary for explosion (Pascanu, Mikolov & Bengio (2013))
- Spectral radius ρ (largest absolute eigenvalue)
  - In a linear RNN, $\rho(W) > 1$ is necessary for exponential explosion
  - $\rho(W) = 1$ with a Jordan block still gives polynomial growth
  - With φ the local Jacobians $D_t W$ matter: $\rho(W) = 0$ yet $\rho(DW) > 1$ is possible
- Stable VAR: impulse responses and gradients decay asymptotically like $\rho(A)^k$
  - At finite horizons $\|A^k\|_2$ can exceed 1: $A = \left(\begin{smallmatrix} 0.5 & 4 \\ 0 & 0.5 \end{smallmatrix}\right)$ has $\rho(A) = 0.5$, yet $\|A^2\|_2 \approx 4.0$
  - Up to a factor k^{m-1} if A has a Jordan block of size m
  - Jordan block: a repeated eigenvalue without a full set of eigenvectors
- DUx_t is an innovation only if the inputs x_t are centred white noise
  - Long-memory inputs would pass through the network

Back

Geometric versus Hyperbolic Decay, Fitted

- Constant gates, i.i.d. candidate input g_t, stationary start: $\mathrm{Corr}(c_t, c_{t-k}) = f^k$
- ARFIMA with $0 < d < 1/2$: $\rho_k \sim Ck^{2d-1}$
  - ARFIMA: autoregressive fractionally integrated moving average
- S&P 500 absolute returns since 2000: ACF (AutoCorrelation Function) 0.289 at lag 1, 0.132 at lag 50
  - Geometric decay with its own amplitude, $\rho_k = \rho_1 f^{k-1}$ (an AR(1) signal observed with noise), matches both lags with $f = 0.984$
  - $b_f = 4.1$, half-life 43 days
  - That f implies 0.027 at lag 150; the data show 0.059, more than twice as much
- A constant gate $f < 1$ is a short-memory filter
  - A long context or explicit long averages, as in HAR (Test 2), only approximate long memory over a finite range of lags

Back

Other Foundation Models and Their Corpora (1/2)

- Chronos-2: training corpus of selected Chronos and GIFT-Eval datasets
  - Plus synthetic multivariate series
- Benchmarks: GIFT-Eval (General Time Series Forecasting Model Evaluation, benchmark GIFT-Eval) (Aksu et al. (2024)) and fev-bench
- TimesFM corpus: Google Trends, Wikipedia page views, synthetic series
  - Also the M4 forecasting-competition series, electricity, traffic, weather
- Release dates of the weights, from the Hugging Face repositories
  - TimesFM-2.5: updated weight version announced on 15.09.2025, released on 01.10.2025
  - Chronos-2: announced on 20.10.2025, released on 30.10.2025
- Checkpoints used: the revision pinned in the code and the first 8 characters of the SHA-256 of the weights file
  - Chronos-2: `ddcda3c7`, in the repository since 30.10.2025
  - Chronos-Bolt (small): `06a6a19b`, since 13.11.2024
  - TimesFM-2.5: `2f776efe`, since 01.10.2025

Back

Other Foundation Models and Their Corpora (2/2)

- **Moirai** (Woo et al. (2024)): one Transformer for any frequency and any number of variables
- The output is a mixture of distributions

$$p(y) = \sum_{k=1}^{4} w_k \, p_k(y), \qquad w_k \geq 0, \quad \sum_{k=1}^{4} w_k = 1$$

  - p_k: Student-t, log-normal, negative binomial and low-variance Normal densities
  - The weights w_k and the component parameters are network outputs, so the shape adapts to the series
- Moirai 2.0: Liu et al. (2025)
- **Lag-Llama** (Rasul et al. (2023)): decoder-only model with lagged values as features and a Student-t output

Back

A Formal Contamination Protocol (1/2)

- Pre/post regression on the loss differential $d_t = L_t^{FM} - L_t^{bench}$, with HAC standard errors

$$d_t = \beta_0 + \beta_1 1\{t \geq \tau\} + u_t$$

  - L_t^{FM}, L_t^{bench}: losses of the foundation model and of the benchmark on day t
  - τ: release date of the weights (a common cutoff when several models are compared)
  - Memorisation predicts $\beta_1 > 0$; H_0: $\beta_1 = 0$

- Report the minimum detectable effect (MDE) of β_1 (AI section)
  - Its floor 0.0157 uses the long-run variance of d_t in each period
  - The pre-release days, which include the April 2025 shock, are about 2.9 times as variable as the later ones
  - With one common variance the floor would be 0.0111

Back

A Formal Contamination Protocol (2/2)

- Pre-registration: forecasts submitted before the outcome exists make contamination impossible by design
  - Live benchmarks: TS-Arena (Time Series, in the name TS-Arena) (Meyer et al. (2025)), Impermanent (Garza et al. (2026))
  - Our post-release window is a manual version of the same idea
- Ablations: removing the language-model part of LLM-based forecasters does not degrade accuracy (Tan et al. (2024))
  - It often improves it
- Electricity prices: Pan & Ezzat (2026) benchmark on two datasets to limit contamination
- Result: foundation models are competitive but do not consistently beat domain-specific methods

Back

Return Forecasts, Full Table

R^2 (%) [DM p]	Historical mean	AR(1)	LSTM (5 seeds)	Chronos-2	Chronos-Bolt	TimesFM-2.5
S&P 500	+0.15 [0.095]	+0.94 [0.639]	+0.93 [0.090]	−0.26 [0.418]	−1.57 [0.078]	−3.35 [0.001]
Bitcoin	+0.07 [0.691]	−0.28 [0.326]	+0.04 [0.243]	−0.35 [0.413]	−0.61 [0.070]	−2.29 [0.001]
BET	+0.30 [0.081]	−0.00 [1.000]	+1.15 [0.014]	−0.39 [0.562]	+0.12 [0.831]	−3.27 [0.070]

- Point forecasts: grid medians (grid means: Design of the Return Test)
- DM: two-sided test on squared errors against the zero forecast, Newey–West variance

Back

Nested Forecasts, the BET AR(1) and Rolling Windows

- BET AR(1), $R^2 = -0.00\%$ and DM $p = 1.000$, is not a degenerate forecast
  - S.d. 0.074%, average slope 0.046
  - Squared-error gains of -12.7 in 2020 and $+14.7$ in 2026 cancel
  - Without these two years $R^2 = -0.11\%$
- Rolling or re-trained models: Giacomini & White (2006) test the forecasting method, estimation window included
  - Diebold (2015) on the use and abuse of DM
- SPA refines White's Reality Check (White (2000))
  - Very poor models (TimesFM) no longer dilute the test
  - Both use a stationary bootstrap with mean block 20 days

Back

Multiple Testing (1/2): Holm

- **Holm**: order the M p-values, $p_{(1)} \leq \cdots \leq p_{(M)}$
  - Reject while $p_{(j)} < 0.05/(M - j + 1)$, stop at the first non-rejection
- The adjusted $p_{(j)}$ is $\max_{i \leq j} \min\{1, (M - i + 1)\, p_{(i)}\}$
  - Toy example, $M = 3$, $p = 0.01, 0.02, 0.04$
  - Thresholds $0.05/3 = 0.0167$, $0.05/2 = 0.025$, 0.05: all three rejected
  - Bonferroni (0.0167 for all) rejects only the first
  - Adjusted p: 0.03, 0.04, 0.04
  - Chapter: the BET LSTM has the smallest of the 18 one-sided DM p-values, 0.0072
  - So its adjusted p is $18 \times 0.0072 = 0.130$

Back

Multiple Testing (2/2): Romano–Wolf and SPA

- **Romano–Wolf**: statistics $t_k = \sqrt{T}\,\bar{d}_k/\hat{\omega}_k$, ordered $t_{(\mathbf{1})} \geq \cdots \geq t_{(K)}$

$$p^{RW}_{(j)} = \max_{i \leq j} P^*\Big(\max_{k \in \mathcal{A}_i} t^*_k \geq t_{(i)} \Big), \qquad t^*_k = \frac{\sqrt{T}\,(\bar{d}^*_k - \bar{d}_k)}{\hat{\omega}_k}$$

  - With $\bar{d}_k$ the mean loss gain of model k over the zero forecast and $\hat{\omega}_k$ its HAC standard deviation
  - $\mathcal{A}_i$: the hypotheses not yet rejected at step i, i.e. $(i), \ldots, (K)$
  - P^*, $\bar{d}^*_k$: probability and mean under the stationary bootstrap
  - The threshold comes from the maximum of correlated statistics, so the procedure is less conservative than Holm for forecasts of one market
- **SPA**: one joint test of H_0: no model has a lower expected loss than the benchmark
  - Statistic $\max\left(0, \max_k t_k\right)$
  - p-value from a stationary bootstrap
  - Much worse models ($t_k \leq -\sqrt{2 \ln \ln T}$) are not recentred at zero, so they do not dilute the test
- Which question each answers
  - Holm and Romano–Wolf: *which* comparisons are significant with at most a 5% chance of any false rejection
  - SPA: *whether any* model beats the benchmark

Back

QLIKE, DM and MCS, Full Table

Model	QLIKE	DM vs log-HAR	p	MCS p	QLIKE after 3.11.2025
Random walk	0.313	+4.44	<0.001	0.005	0.356
HAR	0.238	+1.17	0.241	0.33	0.251
log-HAR	0.226	–	–	0.95	0.223
GARCH-t	0.273	+3.45	<0.001	0.03	0.279
LSTM	0.275	+1.66	0.097	0.29	0.217
Chronos-2	0.223	−0.61	0.539	1.00	0.226
Chronos-Bolt	0.224	−0.20	0.842	0.95	0.219
TimesFM-2.5	0.225	−0.07	0.948	0.95	0.226

- Literature: Brini (2026) compares nine zero-shot foundation models with eight HAR-type models on 50 assets
  - Pooled losses favour foundation models because of a few outlier assets
  - While asset by asset only one small model beats log-HAR at every horizon, narrowly
- Goel et al. (2025): TimesFM zero-shot is a reasonable baseline
  - Incremental fine-tuning is needed to beat econometric models

Back

VaR 1% Breaches and Kupiec Tests, Full Table

Breaches [Kupiec p]	HS	GARCH-t	FHS	Chronos-2 quantiles	FHS + Chronos-2 vol.	FHS + TimesFM vol.
S&P 500 (26.9)	37 [0.065]	44 [0.002]	30 [0.559]	45 [0.001]	25 [0.705]	27 [0.989]
Bitcoin (28.2)	19 [0.065]	27 [0.822]	25 [0.539]	33 [0.374]	27 [0.822]	27 [0.822]
BET (26.8)	32 [0.322]	41 [0.010]	30 [0.536]	40 [0.016]	26 [0.884]	32 [0.322]

- In brackets after the market: the expected number of breaches, 1% of the days
- Kupiec: likelihood-ratio test of a breach rate $\alpha = 1\%$, with x breaches in T days and $\hat{\pi} = x/T$

$$\mathrm{LR}_{uc} = -2 \ln \frac{(1-\alpha)^{T-x}\alpha^{x}}{(1-\hat{\pi})^{T-x}\hat{\pi}^{x}} \sim \chi^{2}_{1}$$

Back

COVID-19 Crash, Scaling and Overshoot

- Chronos-2 standardises the context by its mean and standard deviation, then applies $\sinh^{-1}$ (Ansari et al. (2025))
  - On 16.03.2020 the crash raised the 512-day standard deviation by 9.7%
  - Mean absolute scale of Chronos: 3.2% (Seminar 14, exercise A6)
  - So the widening mixes the scale change and the reading of the crash days
  - Only an ablation could separate them
- Not a unit error: after the crash day Chronos-2 forecast a mean $|r|$ of 12.5% for the next day
  - Then FHS multiplies it by the 1% quantile of $r/\hat{\sigma}$ (4.7)
  - The filter extrapolates a one-day spike as a new level
  - VaR 4.6 times the largest loss, a high capital cost that the FZ0 loss penalises

Back

Berkowitz Tests and Quantile Scores, Three Markets

Market	Berkowitz p: Chronos-2	FHS	Below $q_{1\%}$ (C2)	QS: C2 / FHS	DM [p]
S&P 500	0.019	0.599	1.67%	0.510 / 0.506	+3.06 [0.002]
Bitcoin	0.595	0.306	1.17%	1.526 / 1.529	−0.90 [0.369]
BET	0.045	0.436	1.50%	0.465 / 0.460	+3.48 [<0.001]

- The Berkowitz test evaluates the linearly interpolated grid distribution function
  - And only μ, σ, not serial independence of the PIT
- QS weights the 21 levels equally
  - The CRPS would weight each level by its spacing
  - C2: Chronos-2; on Bitcoin no difference is detected between Chronos-2 and FHS

Back

FZ0 Losses and MCS p-Values, Full Table

FZ0 [MCS p]	HS	GARCH-t	FHS	Chronos-2 quantiles	FHS + Chronos-2 vol.	FHS + TimesFM vol.
S&P 500	1.309 [0.14]	1.026 [0.19]	0.999 [1.00]	1.147 [0.14]	1.084 [0.14]	1.059 [0.18]
Bitcoin	2.237 [0.34]	2.192 [0.60]	2.192 [0.60]	2.181 [0.64]	2.151 [1.00]	2.160 [0.70]
BET	1.230 [0.015]	1.001 [0.34]	0.963 [1.00]	1.032 [0.24]	1.007 [0.34]	1.027 [0.20]

- Bitcoin: the hybrid FHS with Chronos-2 volatility has the lowest FZ0 loss
  - But the differences are not significant
- FZ0 for (VaR 2.5%, ES 2.5%)
  - The MCS uses the $T_{\max}$ statistic and a block bootstrap

Back

Conformal Recalibration, Rules and Results

After 250 calibration days: breaches [cc p / DQ p]	FHS	Chronos-2 raw	FHS + C2 vol.	Chronos-2 + ACI
S&P 500 (24.4)	28 [0.097 / 0.014]	42 [0.002 / <0.001]	23 [0.058 / <0.001]	21 [0.307 / 0.049]
Bitcoin (25.7)	24 [0.449 / 0.108]	31 [0.410 / 0.026]	26 [0.087 / 0.004]	24 [0.449 / 0.191]
BET (24.2)	26 [0.709 / 0.727]	36 [0.048 / 0.184]	23 [0.776 / 0.011]	25 [0.762 / 0.032]
FZ0, S&P 500 [DM p]	1.013	1.142 [0.009]	1.087 [0.022]	1.205 [0.001]

- $\hat{Q}$: the $\lceil (1-\alpha_t)(250+1) \rceil$-th smallest past score ($\lceil \cdot \rceil$: rounding up)
  - $+\infty$ if that rank exceeds 250 or $\alpha_t \leq 0$, $-\infty$ if $\alpha_t \geq 1$
  - With $\gamma = 0.005$: infinite VaR on 24 (S&P 500), 100 (Bitcoin), 270 (BET) days
- Guarantee with $T = 2\,443$, $\gamma = 0.001$: bound 0.41, uninformative for a 1% rate
  - Without exchangeability, weighted conformal bounds the gap by the drift (Barber et al. (2023))
- ES can be evaluated only jointly with VaR, through losses consistent for the pair such as FZ0 (Nolde & Ziegel (2017))
  - ES comparisons face finite-sample precision limits (Pele & Mazurencu-Marinescu-Pele (2026))
  - A GPD (Generalised Pareto Distribution) tail (McNeil & Frey (2000)) is an alternative for ES

Back

Case Study, Replication with Course Data

- Monthly SPY total returns from 1993 (course data) in excess of the 3-month Treasury bill rate
  - Treasury bill series: FRED `TB3MS`
  - Predictors of Welch & Goyal (2008) available: Treasury-bill rate (`tbl`), long-term yield (`lty`), term spread (`tms`), default yield spread (`dfy`), inflation (`infl`), all from FRED; stock variance (`svar`) from daily S&P 500 returns; the lagged excess return: 7 of the 15
  - Data limitation: the valuation ratios (dividend-price `dp`, dividend yield `dy`, earnings-price `ep`, payout `de`, book-to-market `b/m`), net equity expansion (`ntis`) and the bond returns (long-term return `ltr`, default return spread `dfr`) are not in the course data; about 30 years instead of 90
  - Method unchanged: $\gamma = 2$, $T = 12$, $P \leq 12\,000$, $\log_{\mathbf{10}} z = -3, \ldots, 3$, $1\,000$ draws
- Course-data results
  - Linear kitchen sink, $z = 0^+$ ($c = 7/12$): $R^{\mathbf{2}} = -847\%$, Sharpe -0.03
  - Linear, $z = 10^{\mathbf{3}}$: $R^{\mathbf{2}} = 0.26\%$, information ratio 0.25 ($t = 1.3$) against SPY
  - RFF, $c = 1\,000$, $z = 10^{\mathbf{3}}$: $R^{\mathbf{2}} = -5.7\%$, information ratio -0.12 ($t = -0.6$) against SPY and 0.07 ($t = 0.4$) against the linear model

Back

Monthly Alignment and Missing Inflation

- Each predictor month forecasts the next calendar-month return
- A missing predictor does not move its target to a later month
  - Lags and targets keep the monthly calendar
- Inflation is the value known at the end of each month: the ALFRED (ArchivaL Federal Reserve Economic Data) vintage of CPIAUCSL (FRED series: Consumer Price Index for All Urban Consumers, all items) on that day, without later revisions
  - Today's revised series would add look-ahead (monthly correlation with the real-time series: 0.94)
  - October 2025 was never published: in December 2025 November inflation cannot be computed and is not imputed
- The test targets run from February 1997 to August 2026: 354 forecasts
  - Except January 2026
- Volatility scaling uses 12 calendar months, before incomplete rows are removed
- The regression uses the latest 12 complete observations
  - Around missing data this window spans more than 12 months

Back

The Discovery Loop, Reusable Prompts (1/2)

- **Literature**: `"Find evaluations of time-series foundation models on financial returns or volatility since 2024: models, data, loss, benchmark, and whether the test period overlaps the pretraining corpus."`
  - Semantic Scholar, Google Scholar, Elicit help with the search
- **Hypotheses**: `"List predictions that separate memorisation from skill for a zero-shot volatility forecaster: pre vs post release, never-public series, simulated series with the same dynamics."`
  - Diagnostics, not proof: a simulated log-HAR path tests transfer to that process
  - A pre/post gap can also reflect a regime change
- **Data and code**: `"From the daily SPY realised-variance forecasts of Chronos-2 and log-HAR, compute the QLIKE differential Chronos-2 minus log-HAR, test pre vs post 3 November 2025 with HAC errors, and report the minimum detectable effect at 80% power."`
  - Any assistant (Claude, ChatGPT, Gemini, GitHub Copilot)

Back

The Discovery Loop, Reusable Prompts (2/2)

- **Robustness**: `"Design a pre-registered test with more power: which cross-section of series, which variance proxy, which panel DM test, and how many post-release days."`
- **Critique**: `"Attack this claim: no memorisation, because the post-release DM test is insignificant. What is wrong with the inference?"`
  - Expect: low power, absence of evidence, and different regimes
  - The April 2025 tariff shock lies in the pre-release sample, the post-release window is calmer
- **Report**
  - Model versions and weight release dates
  - Context length, mean or median
  - Every seed and window tried

Back

Further Reading

- Gu, Kelly & Xiu (2020): machine learning for the cross-section of expected returns
- Kelly, Malamud & Zhou (2024): why very large models can predict returns better
- Ansari et al. (2024): Chronos, time series read as a language
- Crypto applications
  - Liquidity forecasting Agakishiev, Härdle & Pele (2025)
  - Early warning for "zombie" crypto-assets Będowska-Sójka, Wójcik & Pele (2026)
  - Statistical classification of crypto-assets Pele et al. (2023)
- Textbooks: Franke, Härdle & Hafner (2019) and Härdle, Chen & Overbeck (2017)
  - For the statistics of financial markets behind the benchmarks

Back

Acronyms used in this material (1/2)

- **ACF**: AutoCorrelation Function
- **ACI**: Adaptive Conformal Inference
- **AI**: Artificial Intelligence
- **AR**: AutoRegressive model
- **ARFIMA**: AutoRegressive Fractionally Integrated Moving Average (model)
- **BET**: Bucharest Exchange Trading (index)
- **CI**: Confidence Interval
- **COVID-19**: COronaVIrus Disease 2019
- **CPU**: Central Processing Unit
- **CRPS**: Continuous Ranked Probability Score
- **CRSP**: Center for Research in Security Prices
- **DM**: Diebold–Mariano (test)
- **DQ**: Dynamic Quantile (test)
- **ES**: Expected Shortfall
- **ETF**: Exchange-Traded Fund
- **EWMA**: Exponentially Weighted Moving Average
- **FHS**: Filtered Historical Simulation
- **FM**: Foundation Model
- **FRED**: Federal Reserve Economic Data
- **FRTB**: Fundamental Review of the Trading Book
- **FZ0**: Fissler–Ziegel loss, zero-homogeneous version
- **GARCH**: Generalised AutoRegressive Conditional Heteroskedasticity
- **GIFT**: General Time Series Forecasting Model Evaluation (benchmark GIFT-Eval)
- **GPD**: Generalised Pareto Distribution
- **GRU**: Gated Recurrent Unit
- **GW**: Giacomini–White (test)
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors)
- **HAR**: Heterogeneous AutoRegressive (model)

Acronyms used in this material (2/2)

- **HS**: Historical Simulation
- **i.i.d.**: independent and identically distributed
- **IR**: Information Ratio
- **LASSO**: Least Absolute Shrinkage and Selection Operator
- **LLM**: Large Language Model
- **LN**: Layer Normalisation
- **LR**: Likelihood Ratio (test)
- **LSTM**: Long Short-Term Memory (network)
- **M4**: the fourth Makridakis forecasting competition
- **M6**: the sixth Makridakis forecasting competition
- **MCS**: Model Confidence Set
- **MDE**: Minimum Detectable Effect
- **MHA**: Multi-Head Attention
- **MLP**: MultiLayer Perceptron
- **MSE**: Mean Squared Error
- **NBER**: National Bureau of Economic Research
- **PIT**: Probability Integral Transform
- **QLIKE**: Quasi-LIKElihood (loss function)
- **QS**: Quantile Score
- **ReLU**: Rectified Linear Unit
- **RFF**: Random Fourier Features
- **RNN**: Recurrent Neural Network
- **RoPE**: Rotary Position Embeddings
- **RPS**: Ranked Probability Score
- **RV**: Realised Variance
- **SPA**: Superior Predictive Ability (test)
- **SSE**: Sum of Squared Errors
- **T5**: Text-to-Text Transfer Transformer
- **TS**: Time Series (in the name TS-Arena)
- **US**: United States
- **VAR**: Vector AutoRegression
- **VaR**: Value-at-Risk

References (1/7)

- Agakishiev, I., Härdle, W.K., Pele, D.T. (2025). Cryptocurrency liquidity forecasting. In *Springer Handbooks of Computational Statistics*, 41–60. Springer.
- Aksu, T., Woo, G., Liu, J., Liu, X., Liu, C., Savarese, S., Xiong, C., Sahoo, D. (2024). GIFT-Eval: a benchmark for general time series forecasting model evaluation. arXiv:2410.10393.
- Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2), 579–625.
- Ansari, A.F., Shchur, O., Küken, J., Auer, A., et al. (2025). Chronos-2: from univariate to universal forecasting. arXiv:2510.15821.
- Ansari, A.F., Stella, L., Turkmen, C., Zhang, X., et al. (2024). Chronos: learning the language of time series. *Transactions on Machine Learning Research*; arXiv:2403.07815.
- Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J. (2023). Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2), 816–845.
- Będowska-Sójka, B., Wójcik, P., Pele, D.T. (2026). Early warning systems for cryptocurrency markets: predicting 'zombie' assets using machine learning. *North American Journal of Economics and Finance*, 81, 102543.
- Bengio, Y., Simard, P., Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157–166.
- Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4), 465–474.

References (2/7)

- Brini, A. (2026). Forecasting realized volatility with time series foundation models: a comparison with econometric benchmarks. arXiv:2607.05291.
- Campbell, J.Y., Thompson, S.B. (2008). Predicting excess stock returns out of sample: can anything beat the historical average? *Review of Financial Studies*, 21(4), 1509–1531.
- Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.
- Clark, T.E., West, K.D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291–311.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196.
- Das, A., Kong, W., Sen, R., Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. *Proceedings of ICML*, PMLR 235; arXiv:2310.10688.
- Diebold, F.X. (2015). Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics*, 33(1), 1.
- Diebold, F.X., Gunther, T.A., Tay, A.S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4), 863–883.
- Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

References (3/7)

- Engle, R.F., Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4), 367–381.
- Fischer, T., Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669.
- Fissler, T., Ziegel, J.F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4), 1680–1707.
- Franke, J., Härdle, W.K., Hafner, C.M. (2019). Statistics of Financial Markets: An Introduction, 5th ed. Springer.
- Garza, A., Rosillo, R., Mendoza-Smith, R., Salinas, D., Williams, A.R., Ashok, A. (2026). Impermanent: a live benchmark for temporal generalization in time series forecasting. arXiv:2603.08707.
- Giacomini, R., White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6), 1545–1578.
- Gibbs, I., Candès, E. (2021). Adaptive conformal inference under distribution shift. *NeurIPS*; arXiv:2106.00170.
- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.
- Gneiting, T., Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

References (4/7)

- Goel, A., Pasricha, P., Magris, M., Kanniainen, J. (2025). Foundation time-series AI model for realized volatility forecasting. arXiv:2505.11163.
- Gu, S., Kelly, B., Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.
- Hansen, P.R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4), 365–380.
- Hansen, P.R., Lunde, A., Nason, J.M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- Härdle, W.K., Chen, C.Y.-H., Overbeck, L. (eds.) (2017). Applied Quantitative Finance, 3rd ed. Springer.
- Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Hornik, K., Stinchcombe, M., White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366.
- Kelly, B., Malamud, S., Zhou, K. (2024). The virtue of complexity in return prediction. *Journal of Finance*, 79(1), 459–503.
- Kingma, D.P., Ba, J. (2015). Adam: a method for stochastic optimization. *ICLR*; arXiv:1412.6980.

References (5/7)

- LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Liu, C., Aksu, T., Liu, J., Liu, X., et al. (2025). Moirai 2.0: when less is more for time series forecasting. arXiv:2511.11698.
- Lo, A.W., MacKinlay, A.C. (1990). An econometric analysis of nonsynchronous trading. *Journal of Econometrics*, 45(1–2), 181–211.
- Lopez-Lira, A., Tang, Y., Zhu, M. (2025). The memorization problem: can we trust LLMs' economic forecasts? arXiv:2504.14765.
- Makridakis, S., Spiliotis, E., Hollyman, R., Petropoulos, F., Swanson, N., Gaba, A. (2025). The M6 forecasting competition: bridging the gap between forecasting and investment decisions. *International Journal of Forecasting*, 41(4), 1315–1354.
- McNeil, A.J., Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7(3–4), 271–300.
- Meyer, M., Kaltenpoth, S., Albers, H., Zalipski, K., Müller, O. (2025). TS-Arena: a live forecast pre-registration platform. arXiv:2512.20761.
- Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J. (2023). A time series is worth 64 words: long-term forecasting with Transformers. *ICLR*; arXiv:2211.14730.
- Nolde, N., Ziegel, J.F. (2017). Elicitability and backtesting: perspectives for banking regulation. *Annals of Applied Statistics*, 11(4), 1833–1874.
- Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y. (2020). N-BEATS: neural basis expansion analysis for interpretable time series forecasting. *ICLR*; arXiv:1905.10437.

References (6/7)

- Pan, Z., Ezzat, A.A. (2026). Evaluating time series foundation models for electricity price forecasting: contamination risk, distributional shifts, and covariate dependence. arXiv:2607.02623.
- Pascanu, R., Mikolov, T., Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *ICML*; arXiv:1211.5063.
- Patton, A.J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256.
- Patton, A.J., Ziegel, J.F., Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, 211(2), 388–413.
- Pele, D.T., Mazurencu-Marinescu-Pele, M. (2026). Finite-sample precision limits for expected shortfall forecast comparisons. *Mathematics*, 14(13), 2316.
- Pele, D.T., Wesselhöfft, N., Härdle, W.K., Kolossiatis, M., Yatracos, Y.G. (2023). Are cryptos becoming alternative assets? *The European Journal of Finance*, 29(10), 1064–1105.
- Rahimi, A., Recht, B. (2007). Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20.
- Rahimikia, E., Ni, H., Wang, W. (2025). Re(Visiting) time series foundation models in finance. arXiv:2511.18578.
- Rasul, K., Ashok, A., Williams, A.R., Ghonia, H., et al. (2023). Lag-Llama: towards foundation models for probabilistic time series forecasting. arXiv:2310.08278.

References (7/7)

- Romano, J.P., Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), 1237–1282.
- Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.
- Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T. (2020). DeepAR: probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3), 1181–1191.
- Tan, M., Merrill, M.A., Gupta, V., Althoff, T., Hartvigsen, T. (2024). Are language models actually useful for time series forecasting? *NeurIPS*; arXiv:2406.16964.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. (2017). Attention is all you need. *NeurIPS*; arXiv:1706.03762.
- Welch, I., Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4), 1455–1508.
- White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097–1126.
- Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., Sahoo, D. (2024). Unified training of universal time series forecasting Transformers. *ICML*; arXiv:2402.02592.