








Modelling Financial Markets

Chapter 15: LLMs and Sentiment Analysis



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

Can We Trade the Sentiment That Models Read in the News?

> *"I find that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals, and unusually high or low pessimism predicts high market trading volume."*
>
> — *Tetlock (2007)*

- The chapter compares dictionaries, FinBERT, zero-shot LLMs (large language models) and embeddings on the same labelled texts
  - on Twitter, the best is the simplest supervised model, TF-IDF (weighted word frequencies) + logistic regression: 83.2% accuracy, against 74.2% for Qwen2.5-7B and 72.5% for FinBERT
- On 16 large US (United States) stocks with 106 820 news headlines, the score explains the same-day price move
  - FinBERT: +39.2 bp per standard deviation of the score ($t = +17.27$); bp: basis points, 1 bp = 0.01%
- A daily long-short strategy on the sign of the score brings no net profit
  - Qwen2.5-7B: +5.64 bp per day before costs, $p = 0.071$ after the Holm correction
  - the break-even cost, 1.5 bp per trade, is below the assumed cost of 5 bp
- LLMs remember market history
  - asked only for the month, Qwen2.5-14B separates up from down months of the S&P 500 (Standard & Poor's) with an AUC (Area Under the, ROC Curve) of 0.69 (area under the ROC curve; 0.5 = no discrimination)

▸ Central chart: Cumulative Strategy Returns: No Net Profit

Can Models Read News Sentiment, and Can We Trade It?

From Text to a Trading Signal

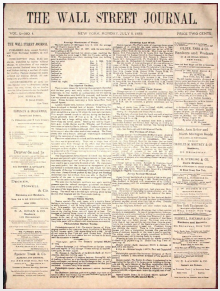
- Part I: text as data, dictionaries, evaluation of classifiers
  - FinBERT and LLMs (large language models), embeddings
- Part II: look-ahead bias, the trading signal, scores as estimates
  - Two case studies and one pass of the research loop assisted by AI (artificial intelligence)
- Detailed examples and advanced topics are in Further Topics, after the Conclusions

What you should be able to do

- compute the sentiment of a text with a dictionary, FinBERT and a zero-shot LLM (Large Language Model)
- interpret the paired test that compares two classifiers on the same texts
- interpret the tests for training overlap and memorised outcomes in an LLM backtest
- interpret the test of a sentiment signal with correct timing, few-cluster inference, costs and a multiplicity correction
- derive the attenuation of the slope when a model score is a generated regressor
- interpret prediction-powered inference intervals
- Further Topics and Appendix: reading for familiarity

Markets Run on Words

- Prices move when information arrives, and most information arrives as text
  - News, company filings, earnings calls, analyst reports, central-bank statements, social media
- Financial newspapers are as old as organised markets
  - The Wall Street Journal: first issue on 8 July 1889
- Today: millions of documents per day, far more than any analyst can read


THE WALL STREET JOURNAL.

The Wall Street Journal, first issue, 8 July 1889

Photo: Dow Jones & Company (1889); public domain; Wikimedia Commons

Text as Data, and What We Mean by Sentiment

- Gentzkow, Kelly & Taddy (2019): a document d becomes numbers x_d (counts, scores, vectors)
  - These numbers feed an ordinary statistical model
  - Map x_d to a quantity of interest (sentiment, topic, risk), then use it in an economic question (returns, volatility)
- **Sentiment of a text**: good, bad or neutral news for the asset it mentions
  - Three classes, or the net score $s = P(\text{positive}) - P(\text{negative}) \in [-1, 1]$
  - Not **investor sentiment**, the mood of investors measured from market data (Baker & Wurgler (2006))
- Surveys for finance: Kearney & Liu (2014), Loughran & McDonald (2016)
- This chapter: the score is an **estimate** of the true tone, with measurement error

Early Evidence and Narratives

- Antweiler & Frank (2004): internet message-board posts about stocks help predict volatility
  - The effect on returns is small
- Tetlock (2007): high media pessimism predicts downward pressure on prices
  - Pessimism measured in a Wall Street Journal column
  - The pressure is followed by a reversal
  - Unusually high or low pessimism also predicts high trading volume
- Tetlock, Saar-Tsechansky & Macskassy (2008): negative words in firm-specific news predict lower earnings and returns
  - Measure: the share of negative words
  - Prices incorporate this information only gradually
- Shiller (2017): popular stories spread like epidemics and move economic decisions
  - They can be measured through text
  - Shiller: Nobel Prize in Economics 2013, with Fama and Hansen; narratives can push prices away from fundamentals (Chapters 2 and 17)



Robert J. Shiller, Stockholm, December 2013

Photo: Bengt Nyman (2013); CC BY 2.0; Wikimedia Commons

Four Ways to Turn Text into Sentiment

Method	How it works	Labels needed	Examples
Dictionary	count listed words	none	GI, LM, VADER
Fine-tuned model	network trained on labelled sentences	thousands	FinBERT
Large language model	instruction in plain language (prompt)	none (zero-shot)	Qwen2.5, GPT
Embeddings + classifier	text vectors into a regression	hundreds to thousands	MiniLM

GI: the Harvard General Inquirer, a general-purpose word list; LM (Loughran-McDonald, financial sentiment dictionary): the Loughran–McDonald financial dictionary; VADER: Valence Aware Dictionary and sEntiment Reasoner, a dictionary with rules built for social media; GPT: Generative Pre-trained Transformer, the language models from OpenAI

- We compare all four on the same labelled data, with 95% confidence intervals (CI)
- Then we test whether any of them makes money on daily stock returns
- A fifth route, text regressions with statistical theory, is not tested here (Appendix)
  - Taddy; Ke, Kelly and Xiu

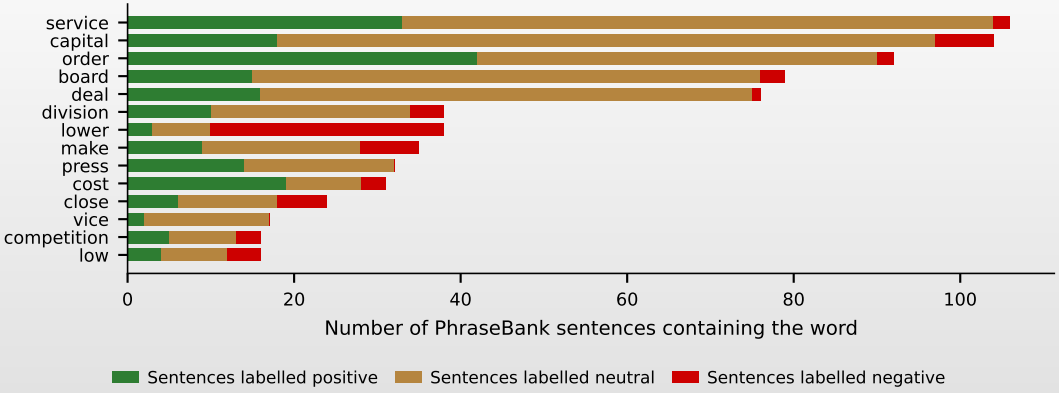
Can a Word List Measure the Tone of a Financial Text?

- **Bag of words**: a document is reduced to its word counts
  - Word order is ignored
  - Tokens: case-insensitive words, without digits and punctuation
- A dictionary lists positive words $\mathcal{P}$ and negative words $\mathcal{N}$
  - P_d, N_d: number of positive and negative words in document d
  - **Tone**: $\tau_d = \dfrac{P_d - N_d}{P_d + N_d} \in [-1, 1]$, and $\tau_d = 0$ if $P_d + N_d = 0$; the class is the sign of τ_d (neutral for $\tau_d = 0$)
- Transparent, fast and free of training data
  - Blind to context and negation

Generic, Financial and Rule-Based Dictionaries

- **Harvard General Inquirer** (GI), word list Harvard IV-4: a psychology dictionary from the 1960s
  - 2 007 negative and 1 637 positive words
- Loughran & McDonald (2011) (LM): built from annual reports (10-K filings) of US firms
  - 2 355 negative and only 354 positive words
  - Their finding: almost three quarters of the GI negative words are not negative in a financial context
  - In the lists used here, 1 552 GI negative words (77%) are not LM negative words
- **VADER** (Hutto & Gilbert (2014)): about 7,500 terms with intensity scores, built for social media
  - Rules for negation ("not good"), intensifiers ("very"), capital letters, exclamation marks and emoticons
  - Compound score in $[-1, 1]$, usual cut-offs ± 0.05
  - The rules fix some context problems, but the vocabulary is generic, like GI

Negative for a Psychologist, Neutral for an Accountant (1/2)



MFM_ch15_dictionaries ▶ Interactive

Negative for a Psychologist, Neutral for an Accountant (2/2)

- Words negative in GI but not in LM
  - Ranked by the number of Financial PhraseBank sentences that contain them
  - Colour: the label human annotators gave the sentence
- "Service" (106 sentences), "capital" (104) and "board" (79) are business nouns, not bad news
  - Not labelled negative: 98%, 93%, 96% of their sentences; "vice" ("vice president"): 100%
- "Lower": only 26% not negative, but "lower costs" is good news
  - The sign depends on the next word, which no word list sees
- Of the 964 sentences whose only GI-negative words are missing from the LM negative list:
  - 89% are not labelled negative

MFM_ch15_dictionaries

Which Texts Do We Judge a Classifier On? Two Labelled Corpora

- Financial PhraseBank (Malo et al. (2014)): 4 846 news sentences
  - Labelled by a pool of 16 annotators with a finance background, 5 to 8 per sentence
  - Question: good, bad or neutral news for the investor?
  - Agreement levels: at least 50%, 66%, 75% or 100% of the sentence's annotators; 2 264 sentences with full agreement, 629 below 66%
  - Labels: 59% neutral, 28% positive, 12% negative
- Twitter Financial News (Hugging Face Hub; MIT licence): headlines from 2022
  - Labelled bearish, bullish or neutral
  - Training set of 9,543, validation set of 2 388 headlines (used for evaluation)
  - Labels: 66% neutral, 20% positive, 15% negative
  - Short, noisy texts with tickers and links: closer to the real news flow
- Two corpora, because a result on one may not carry over to the other

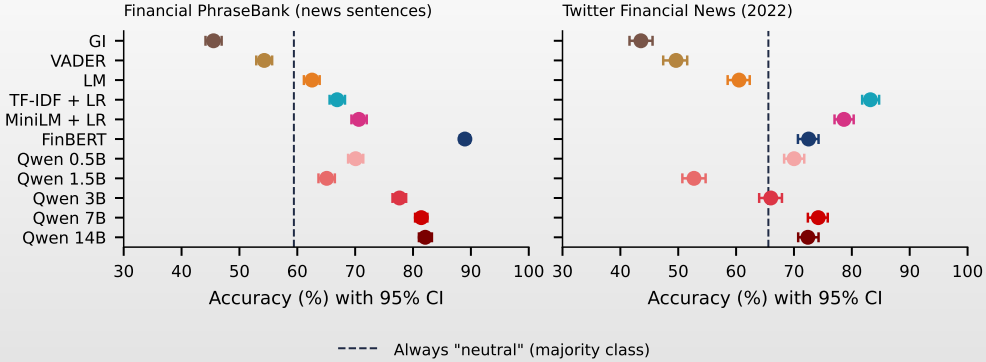
Accuracy or Macro-F1? Metrics for Unbalanced Classes

- The question: how do we compare classifiers when one class dominates?
- Accuracy: $\text{acc} = \frac{1}{n}\sum_{i=1}^{n} 1\{\hat{y}_i = y_i\}$
- For class k: precision $P_k = \dfrac{TP_k}{TP_k + FP_k}$, recall $R_k = \dfrac{TP_k}{TP_k + FN_k}$, $F1_k = \dfrac{2P_kR_k}{P_k + R_k}$
  - TP_k: texts of class k predicted k; FP_k: texts of other classes predicted k; FN_k: texts of class k predicted otherwise
- Macro-F1 $= \frac{1}{K}\sum_{k=1}^{K} F1_k$, with $K = 3$
  - Every class counts the same, however rare (Chapter 13)
- Benchmark "always neutral" on Twitter: 66% accuracy, but a macro-F1 of only 26.4%
  - $F1_{\text{neutral}} = 2 \times 0.656 \times 1/(0.656 + 1) = 0.792$, and $F1 = 0$ for the other two classes, so macro-F1 $= 0.792/3$

When Is One Classifier Better? The McNemar Test on the Same Texts

- Two classifiers A and B on the same n texts: their accuracies are correlated
  - So two separate CIs do not answer the question
- n_{01}: number of texts where A is wrong and B is right
- n_{10}: number of texts where A is right and B is wrong
- Texts on which both classifiers agree carry no information
  - The accuracy gap is $(n_{01} - n_{10})/n$
- Under H_0 (equal accuracy), $n_{01} \sim \text{Binomial}(n_{01} + n_{10}, 0.5)$
  - An exact test (worked example: Appendix)
- For large $n_{01} + n_{10}$: $\chi^2 = (|n_{01} - n_{10}| - 1)^2/(n_{01} + n_{10})$ is approximately χ^2_1
  - -1: continuity correction
- Example: A is wrong and B right on 4 texts, the reverse on 1
  - So $p = 2\,P(X \geq 4) = 2 \times 6/32 = 0.375$, with $X \sim \text{Binomial}(5, 0.5)$

Accuracy of Eleven Classifiers on Two Data Sets (1/2)



TF-IDF: term frequency–inverse document frequency, word counts weighted down for common words; LR: logistic regression

MFM_ch15_classification ▶ Interactive

Accuracy of Eleven Classifiers on Two Data Sets (2/2)

- TF-IDF + LR and MiniLM + LR: the only methods trained on Twitter labels
- All three dictionaries lose to the majority class on Twitter
  - Twitter: GI 43.6%, VADER 49.6%, LM 60.5%; "always neutral": 66%
- FinBERT: 89.0% on PhraseBank, 72.5% on Twitter; Qwen2.5-7B zero-shot: 81.4% and 74.2%
- Best on Twitter, 83.2%: the simplest supervised model, TF-IDF + logistic regression
- Lesson: labelled data from the target domain beat model size
- Type of evidence: comparison of procedures (it shows which method works better on the same data, not why)

MFM_ch15_classification

Does Context Help? From Words to BERT

- A dictionary gives "lower" one fixed sign
  - A language model reads the sentence around it
  - Each token gets a vector that depends on all other tokens (attention, Chapter 14; Vaswani et al. (2017))
- **BERT** (Bidirectional Encoder Representations from Transformers; Devlin et al. (2019))
  - A Transformer encoder pretrained on books and Wikipedia by guessing words hidden at random
  - About 110 million parameters
  - For classification it uses the final vector $h_{\mathbf{CLS}} \in \mathbb{R}^{\mathbf{768}}$ of the special token [CLS (classification token of BERT)] (for *classification*), placed at the start of every text
- Fine-tuning: a classification layer on top of $h_{\mathbf{CLS}}$
  - The whole network is trained on labelled sentences
  - $p = \text{softmax}(Wh_{\mathbf{CLS}} + c)$, with $W \in \mathbb{R}^{\mathbf{3}\times\mathbf{768}}$: probabilities of the negative, neutral and positive classes
  - Training minimises the cross-entropy $-\frac{\mathbf{1}}{n}\sum_i \ln p_{i,y_i}$, with y_i the human label of sentence i

FinBERT: BERT Adapted to Financial Text

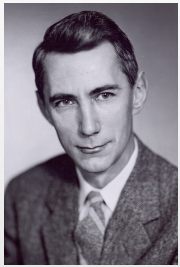
- Araci (2019): FinBERT is BERT further pretrained on financial news, then fine-tuned for sentiment
  - The public model used here (ProsusAI/finbert) was fine-tuned on Financial PhraseBank
  - A few milliseconds per sentence on a laptop
- The score used in this chapter: $s = p_{\text{pos}} - p_{\text{neg}} \in [-1, 1]$
  - The label is $\arg\max_k p_k$
- Why FinBERT as the benchmark: public, fast and widely used in the finance literature
- Huang, Wang & Yang (2023): a FinBERT trained on analyst reports, earnings-call transcripts and 10-K filings
  - It outperforms dictionaries, mainly on sentences that dictionaries misread

In-Sample Evaluation Bias

- FinBERT: 89.0% on PhraseBank, most of which it was trained on
  - Araci (2019): 3,101 of the 4,845 sentences in his version (4 846 here)
  - 72.5% on Twitter: a drop of 16.4 points (from unrounded values)
  - Qwen2.5-7B, not fine-tuned on these labels, also drops by 7.2 points: Twitter headlines are harder
- Difference in differences: $\Delta = (a_{F,P} - a_{F,T}) - (a_{Q,P} - a_{Q,T}) = 9.2$ points
  - Bootstrap CI 95% [6.7; 11.9]
  - $a_{m,D}$: accuracy of model m (F: FinBERT, Q: Qwen) on data D (P: PhraseBank, T: Twitter)
  - Identifying assumption: without training overlap, both models would lose the same amount from PhraseBank to Twitter
  - The CI measures sampling noise, not the assumption (the two architectures may react differently to a new domain)
- Type of evidence: a descriptive difference between the accuracy drops
  - Attributing Δ to training overlap also needs an uncontaminated comparator and no other differential change
  - Qwen was not fine-tuned on PhraseBank, but the texts may be in its pretraining data
- Rule: evaluate a language model only on texts it has never seen
  - As in the out-of-sample tests of Chapters 8 and 13

From Next-Word Prediction to Instructions

- Shannon (1951): English is highly predictable from its past letters
  - A large language model (LLM) estimates $P(w_{t+1} \mid w_1, \ldots, w_t)$ over tokens
  - w_t: the t-th token, a word or part of a word
  - Brown et al. (2020): at scale, tasks are solved from a few examples in the prompt (few-shot)
  - Ouyang et al. (2022): instruction tuning and RLHF (reinforcement learning from human feedback) teach the model to follow requests
- BloombergGPT (Wu et al. (2023)): 50 billion parameters, financial and general text
  - Background: Quantinar: The Next Word Problem, Quantinar: Natural Language Processing



Claude E. Shannon, founder of information theory

Photo: Tekniska museet, Stockholm (undated); CC BY 2.0; Wikimedia Commons

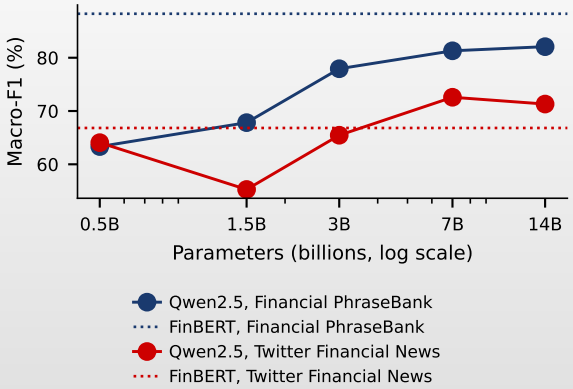
Zero-Shot Classification with Qwen2.5: the Prompt and the Probabilities

- Prompt: a system instruction plus the text to classify
  - `You classify the sentiment of financial news for investors. Answer with one word: positive, negative or neutral.`
  - `Headline: <text> Sentiment:`
- We read the logits z_k (the model's unnormalised scores) of the tokens "positive", "negative", "neutral" at the first generated position
  - $p_k = e^{z_k} / \sum_j e^{z_j}$, normalised over the three answer words
  - Label: $\arg\max_k p_k$
  - Score: $p_{\mathbf{pos}} - p_{\mathbf{neg}}$, as for FinBERT

Temperature and the Qwen2.5 Models

- When generating, the model draws the next token from $p_k(T) = e^{z_k/T} / \sum_j e^{z_j/T}$
  - With temperature $T > 0$
  - $T \to 0$: always the most likely token (greedy decoding), so the same answer on every run; Lopez-Lira & Tang (2026) use temperature 0
  - $T = 1$: the model's own distribution; $T > 1$ flattens the probabilities, so answers vary more
- No text is generated here: the label $\arg\max_k p_k$ is the $T \to 0$ answer
  - The score keeps the $T = 1$ probabilities, with nothing to parse
- Qwen Team (2024): open-weight models, from 0.5 to 72 billion parameters
  - Licences: Apache 2.0; Qwen Research License for 3B
  - Released on 19 September 2024
- We use 0.5B, 1.5B, 3B, 7B and 14B (B: billion parameters)
  - Processing times: Appendix
- Open weights: results can be reproduced
  - The release date bounds what the model could know

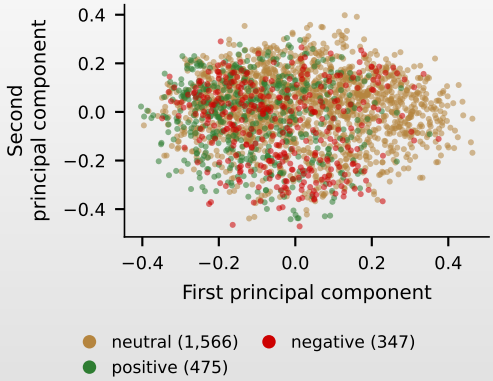
Model Size and Zero-Shot Accuracy



- Macro-F1 of zero-shot Qwen2.5 with the same prompt; dotted lines: FinBERT
- PhraseBank: macro-F1 rises from 63.3% (0.5B) to 82.1% (14B), still below FinBERT (88.3%)
  - FinBERT's lead there is partly training overlap (In-Sample Evaluation Bias)
- Twitter: not monotone
  - 1.5B is the worst (55.3%)
  - 7B the best (72.6%), above FinBERT (66.8%)
  - 7B against 1.5B on accuracy: +21.5 points, McNemar $p < 0.001$
- Small models react strongly to wording
  - The 1.5B model rarely answers "neutral" with this prompt

MFM_ch15_llm_zero_shot ▶ Interactive

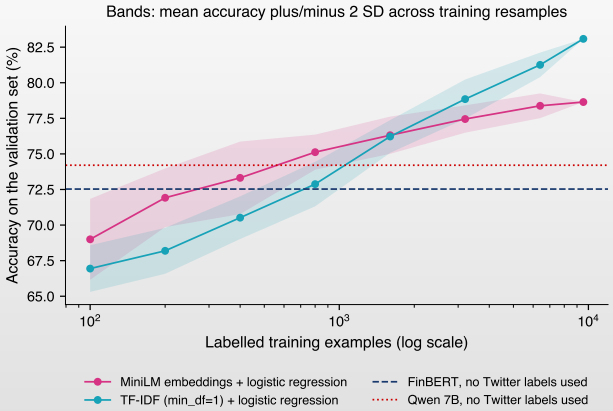
Can Generic Text Vectors Measure Sentiment?



- **Embedding**: a vector $e_d \in \mathbb{R}^k$, close for similar texts
  - word2vec (Mikolov et al. (2013)), Sentence-BERT (Reimers & Gurevych (2019))
  - Here MiniLM (Wang et al. (2020)), $k = 384$, unit length: similarity $\cos(e_a, e_b) = e_a^\top e_b$; embeddings enter a classifier or regression like any feature
- Chart: Twitter validation headlines, first two principal components of the MiniLM vectors
  - Coloured by label
- The first two components do not separate sentiment
  - PCA (principal component analysis, Chapter 3) keeps the directions of largest variance
  - Not those tied to the label
- A logistic regression on all 384 dimensions finds the sentiment direction: 78.6% accuracy on Twitter

MFM_ch15_embeddings ▶ Interactive

How Many Labels Do You Need? (1/2)



MFM_ch15_embeddings ▶ Interactive

How Many Labels Do You Need? (2/2)

- Logistic regression trained on random subsets of the Twitter training set
  - 10 repetitions per size; band: ± 2 standard deviations
- Embeddings + logistic regression: 69.0% with 100 labels, above FinBERT from 400 labels
  - The pretrained vectors already contain most of what is needed
- TF-IDF starts lower (66.9%) but keeps improving
  - About 83% with all labels, against 78.6% for embeddings
  - With many labels, exact words (“upgrade”, “cuts”, “misses”) carry the signal in headlines
- Practical rule: label a few hundred documents from your own domain
  - Do this before trusting any off-the-shelf model

MFM_ch15_embeddings

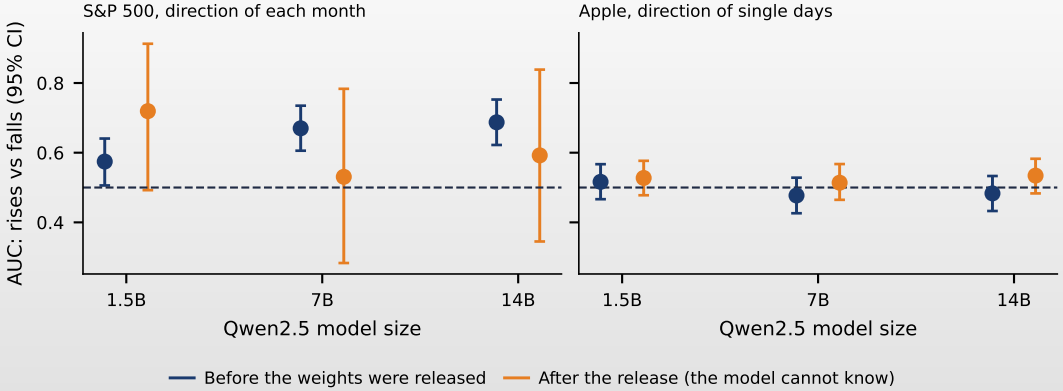
Memorisation of Test Data by Language Models

- An LLM trained on text up to 2024 may have read what happened after a 2015 headline
  - A backtest on 2015 news then mixes forecasting with remembering
- Glasserman & Lin (2023): two biases in sentiment backtests
  - Look-ahead: the model knows the returns that followed the news
  - Distraction: general knowledge about the named company interferes with reading the text; in their tests it mattered more than look-ahead
- Sarkar & Vafa (2024): direct tests show look-ahead in LLM applications to earnings calls and elections

Our Test: Ask About the Past, Without Context

- Monthly: "Did the S&P 500 index rise or fall during <month year>?"
  - Every month from 296 pre-release months plus 23 after
  - Daily: "Did Apple stock close higher or lower on <date> than on the previous trading day?", 500 random days 2010–2023 and 493 days after October 2024
- Score: $P(\text{rise})$ from the answer-word probabilities
  - True direction from the S&P 500 monthly closes and the Apple daily closes
  - AUC (area under the ROC curve; ROC: receiver operating characteristic): probability that a random up period gets a higher score than a random down period; 0.5 = no discrimination (Chapter 13)
- After the release there is nothing to remember
  - If the date carries no predictive information (e.g. seasonality), the population AUC is 0.5
  - The sample AUC varies around it

Memory Measured: Months Yes, Days No (1/2)



MFM_ch15_lookahead ▶ Interactive

Memory Measured: Months Yes, Days No (2/2)

- AUC with bootstrap 95% intervals, months resampled independently
- Months before release: AUC 0.67 (7B, CI [0.61; 0.73]) and 0.69 (14B, CI [0.62; 0.75])
  - Permutation p for 14B <0.001
  - The 1.5B model knows much less (0.57)
- After release (23 months) the intervals include 0.5 but are too wide to prove anything
  - About 101 post-release months would be needed to detect a fall to 0.5 (formal test: Appendix)
- Single days of Apple: AUC between 0.48 and 0.52 before release, no discrimination above chance detected
  - Look-ahead is a danger at the horizon of well-known events, less for daily noise
- Type of evidence: description; the discrimination of months is compatible with memorisation but does not separate it from calendar or regime regularities

MFM_ch15_lookahead

Remedies, Including Point-in-Time Models

- Test only after the model's training cutoff
  - Lopez-Lira & Tang (2026) evaluate GPT-4 (Generative Pre-trained Transformer 4) on post-cutoff headlines
  - Clean, but the window starts late
- Mask company names and dates before scoring (Glasserman & Lin (2023))
  - Then compare slopes with and without masking
  - A diagnostic, since products, events or quotes can still identify the episode
- Point-in-time models, trained only on text available up to date t (Sarkar & Vafa (2024))
  - Drinkall et al. (2024): a series of point-in-time LLMs, each uninformed about facts after its date; models and training data are public
  - He et al. (2025): ChronoBERT and ChronoGPT use only text available at each date; on next-day returns from news their Sharpe ratios are close to a much larger model: look-ahead there is modest
- Report the model version and the release date of its weights with every result

Dated Headlines Matched with Daily Prices

- The question of this section: does headline tone predict returns that we can still trade?
- FNSPID (Financial News and Stock Price Integration Dataset, Dong, Fan & Peng (2024))
  - 15.7 million news records about S&P 500 firms, 1999–2023, from four news websites
  - Creative Commons licence CC BY-NC 4.0 (attribution, non-commercial use)
- Here: 16 large US stocks with daily prices in the course data
  - Technology, banks, GameStop, Strategy
  - 106 820 distinct headlines after removing repeats; 18 282 stock-days with news, 27.10.2009 to 18.12.2023

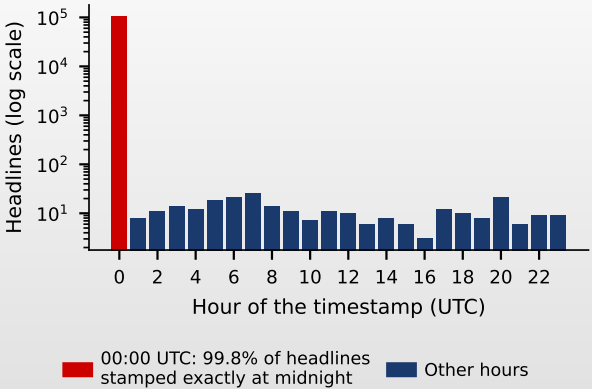


News ticker on the Reuters building, New York

Photo: bgilliard (2005); CC BY-SA 2.0; Wikimedia Commons

▸ Further Topics: headline coverage

News Timestamps: Dates without a Time of Day (1/2)



MFM_ch15_news_data ▶ Interactive

News Timestamps: Dates without a Time of Day (2/2)

- Hour of the FNSPID timestamp in Coordinated Universal Time (UTC), log scale
- 99.8% of headlines are stamped exactly 00:00 UTC
  - The source gives the date, not the time of publication
  - A headline dated d may have appeared before the open, during the session or after the close of day d
- Alignment used here, with no look-ahead
  - Day d: the first trading day on or after the date of the headline
  - Returns on d and $d+1$ may still contain the reaction
  - First safe forecast: the return on $d+2$, close of $d+1$ to close of $d+2$

MFM_ch15_news_data

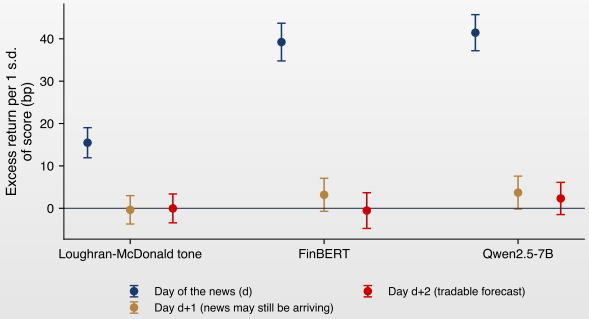
Three Scores and the Excess Return

- The three scores: LM tone, FinBERT net score and Qwen2.5-7B net score with prompt P1 (the one on the zero-shot classification slide)
  - FinBERT net score: $P(\text{pos}) - P(\text{neg})$
  - LM finds a listed word in only 27% of headlines: the rest get tone 0
  - FinBERT labels: 19% negative, 62% neutral, 18% positive
  - Correlations across headlines: FinBERT–Qwen 0.62, FinBERT–LM 0.41, Qwen–LM 0.36
- Excess return: $r^{e}_{i,t} = r_{i,t} - r_{\mathbf{SPY},t}$
  - SPY: an exchange-traded fund tracking the S&P 500 index
  - Market-adjusted: $\beta = 1$ for every stock; the usual alternative is $r_{i,t} - \hat{\alpha}_i - \hat{\beta}_i r_{\mathbf{SPY},t}$, with $\hat{\beta}_i$ estimated before the event

The Regression and Standard Errors Clustered by Day

- Pooled regression over stock-days with news: $r^{e}_{i,d+h} = a + b\, z_{i,d} + \varepsilon_{i,d+h}$, $h = 0, 1, 2$
  - z: daily score standardised over the panel (mean 0, standard deviation 1), so b is the excess return (%) per one standard deviation of score
  - $h = 0$ measures reaction, $h = 2$ measures predictability
- Errors of stocks on the same day are correlated (common market shocks)
  - OLS (ordinary least squares) standard errors (SEs) ignore this: cluster by day (Petersen (2009))
  - $\widehat{V} = (X^\top X)^{-1} \left(\sum_g X_g^\top \hat{\varepsilon}_g \hat{\varepsilon}_g^\top X_g \right) (X^\top X)^{-1}$
  - g: trading days; X: the regressors (constant, z); X_g, $\hat{\varepsilon}_g$: the rows and OLS residuals of day g

Reaction versus Prediction: Headlines Describe Prices



- Slope b in basis points (bp, 1 bp = 0.01%) per one standard deviation of score (s.d. in the chart)
  - 95% CI clustered by day
- Day d: FinBERT +39.2 bp ($t = +17.27$), Qwen +41.4 bp ($t = +19.07$), LM +15.5 bp ($t = +8.53$)
  - Many headlines report the price move itself ("shares jump", "stock falls"): causality runs both ways
- Day $d + 1$: FinBERT +3.2 bp ($t = +1.60$), Qwen +3.7 bp ($t = +1.86$)
  - Part of this is news published after the close of d, not a forecast
- Day $d + 2$: FinBERT −0.5 bp ($t = -0.25$), Qwen +2.3 bp ($t = +1.19$), LM −0.0 bp ($t = -0.01$)

MFM_ch15_news_signal ▶ Interactive

Inference with 16 Stocks: Clustering by Stock and Two-Way

- The question: do the slopes survive if errors are also correlated over time within a stock?
- Firm effects persist, so the errors of one stock are correlated across days: cluster by stock as well
- Two-way clustering (Cameron, Gelbach & Miller (2011)): $\widehat{V}_{\mathbf{2way}} = \widehat{V}_{\mathbf{day}} + \widehat{V}_{\mathbf{stock}} - \widehat{V}_{\mathbf{day} \cap \mathbf{stock}}$
  - Each term is the sandwich formula of the previous slide, with clusters given by day, by stock, and by the stock–day pair
  - A stock–day pair holds a single observation here, so the last term is the White estimator, robust to heteroskedasticity only
  - Subtracting it avoids counting the variance of each observation twice
- Driscoll & Kraay (1998) also allow for correlation across different days

Few Clusters: the Wild Cluster Bootstrap

- With only $G = 16$ stocks, the stock-clustered SE (Standard Error) is too small and the t test rejects too often (Cameron, Gelbach & Miller (2008))
- Wild cluster bootstrap (resampling with random signs per cluster) with the null imposed: fit the restricted model ($b = 0$)
  - Constant $\tilde{a}$ and residuals $\tilde{\varepsilon}_{i,t}$
  - For $j = 1, \ldots, B$, $B = 9{,}999$: Rademacher weights $w_g = \pm 1$ with probability $\frac{1}{2}$, one draw per stock g
  - $r^{e*}_{i,t} = \tilde{a} + w_{g(i)}\,\tilde{\varepsilon}_{i,t}$; re-fit the regression and compute the stock-clustered t^*_j
  - $p = \frac{1}{B}\sum_j 1\{|t^*_j| \geq |t|\}$
- The same sign for all residuals of a stock keeps the correlation within the cluster
- The bootstrap assumes independent stocks, so read it next to the two-way t

Nine Regressions: the Holm Correction

- Nine regressions: 3 scores $\times$ 3 horizons ($h = 0, 1, 2$)
- If no score had any effect, nine tests at 5% would still give $9 \times 0.05 = 0.45$ false rejections on average
- The Holm correction (Chapter 2) keeps the FWER at or below 5%
  - FWER (family-wise error rate): the chance of any false rejection
  - Sort the p-values: $p_{(1)} \leq \cdots \leq p_{(m)}$, with $m = 9$
  - $\tilde{p}_{(k)} = \max_{j \leq k} \min\{1, (m - j + 1)\, p_{(j)}\}$; reject $H_{(k)}$ when $\tilde{p}_{(k)} \leq 0.05$
  - More powerful than Bonferroni ($m\, p_{(k)}$ for all) and valid under any dependence between the tests

Inference with 16 Stocks and Nine Regressions: Results

Score, horizon	b (bp)	t by stock	t two-way	p wild bootstrap	p Holm
FinBERT, d	+39.2	+5.71	+5.67	<0.001	<0.001
Qwen, d	+41.4	+5.10	+5.07	<0.001	<0.001
FinBERT, $d+2$	−0.5	−0.35	−0.30	0.739	1.00
Qwen, $d+2$	+2.3	+2.41	+2.01	0.061	0.37
LM, $d+2$	−0.0	−0.03	−0.02	0.979	1.00

- Qwen at $d+2$: $t = +2.41$ by stock looks significant
  - $p = 0.029$ under a t distribution with 15 degrees of freedom
  - But not after the wild bootstrap and Holm
  - Holm step by step: sort the 9 bootstrap p-values; Qwen at $d+2$ comes 4th, after the three same-day values (all below 0.001), so its Holm $p = (9 - 4 + 1) \times 0.061 \approx 0.37$
  - Rows not shown: LM at d, p <0.001; at $d+1$, $p = 0.272$ (FinBERT), 0.317 (Qwen), 0.802 (LM)
- Type of evidence: association (the variables move together; this does not prove causation)
  - The most positive third of scores already has a cumulative excess return of +0.63% in the 5 days before the news (Further Topics)

A Daily Long-Short Sentiment Strategy

- The question: does the sign of the score earn a profit we can actually collect?
- At the close of $d+1$: for every stock with news and a non-zero score on d, go long or short
  - Long if its score is positive and short if negative
  - Each position hedged with SPY; a one-for-one hedge removes the market move only when $\beta = 1$
  - Equal weights; positions are closed after one day, at the close of $d+2$
  - The LM score is zero on 48% of stock-days: these do not enter the portfolio
- Gross return on $d+2$: $G_{d+2} = \frac{1}{N_d} \sum_{i \in A_d} \text{sign}(s_{i,d}) \left(r_{i,d+2} - r_{\text{SPY},d+2}\right)$
  - $A_d = \{i : |s_{i,d}| > \delta\}$, with threshold $\delta = 0$: the stocks with news and a non-zero score on day d
  - $N_d = |A_d|$: the positions opened at the close of $d+1$; if A_d is empty, $G_{d+2} = 0$ and no costs are paid

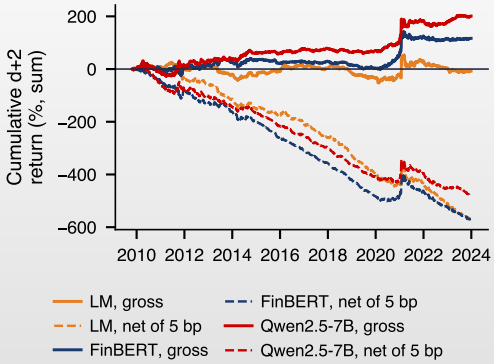
Trading Costs: Round Trips and the Break-Even Cost

- $c = 5$ bp per one-way trade, per unit of traded notional
  - Assumed, not estimated; plausible for large, liquid US stocks
- A round trip (in and out) costs $2c$
- On a day with positions: one round trip for the stock book and one for the SPY hedge
  - Each of notional 1, so $4c = 20$ bp
- Net return: $G^{\text{net}}_{d+2} = G_{d+2} - 4c\,1\{N_d > 0\}$
- Break-even cost c^*: the value of c at which the mean net return is 0
  - I.e. $c^* = \bar{G}_{\text{active}}/4$
  - $\bar{G}_{\text{active}}$: the mean gross return on days with positions

Reading the Strategy Numbers: Sharpe Ratio and Summed Returns

- Inference: Newey–West t (Chapter 2) and block-bootstrap CI for the mean daily return
- Sharpe ratio $SR = \sqrt{252}\,\bar{G}/\hat{\sigma}$ (annualised)
  - $\bar{G}$, $\hat{\sigma}$ are the mean and standard deviation of the daily return G_t over all 3 568 trading days
  - Days without positions enter as 0
  - After costs, G_t is replaced by the net return
  - Example (Qwen2.5-7B): mean $+5.64$ bp and $SR = +0.63$ imply $\hat{\sigma} = \sqrt{252}\times$ mean$/SR \approx 143$ bp per day
  - After costs: positions on 95% of days, so the mean falls by $20\times$ 0.95 bp, to -13.36 bp per day
- The next chart sums daily simple excess returns
  - An additive profit measure, not compounded wealth

Cumulative Strategy Returns: No Net Profit



- Sum of daily excess returns (%), before costs and with 5 bp per trade
- FinBERT: +3.27 bp per day ($t = +1.27$), not different from zero
  - CI 95% [−1.51; +8.93]; LM: −0.21 bp ($t = -0.07$)
- Qwen2.5-7B: +5.64 bp ($t = +2.26$, $p = 0.024$)
  - CI 95% [+0.95; +10.61]; but it is the best of the three scores tested: Holm p across the three = 0.071, not significant
  - Break-even cost +1.5 bp per trade, well below 5 bp: Sharpe +0.63 before, −1.48 after costs
- Type of evidence: association
  - Only 2020–2021 has large gains, not significant at 5% (Further Topics)

MFM_ch15_news_signal ▶ Interactive

The Strategy: the Non-Tradable Day and Memorisation

- Non-tradable day $d+1$: FinBERT -0.22 bp ($t = -0.10$), Qwen $+4.06$ bp ($t = +1.36$)
  - No significant gain even there: almost all of the reaction happens on day d
- Part of the Qwen edge may be memorised: the Qwen2.5 weights were released on 19 September 2024, after the 2010–2023 headlines
  - A headline that names the episode can trigger recall, unlike the Apple test, which gave only the date

MFM_ch15_news_signal

Sentiment as a Generated Regressor: Classical Error

- The question of this section: what happens to the slope and its interval
  - When the regressor is a model output, not the true tone?
- Model: $r = a + \beta s^* + \varepsilon$ with true tone s^*
  - We observe a model output $\hat{s} = s^* + u$
  - Classical error (u independent of s^* and ε): plim $\hat{b} = \beta\lambda$, $\lambda = \mathrm{Var}(s^*)/[\mathrm{Var}(s^*) + \mathrm{Var}(u)] < 1$
  - λ: the reliability of the score, the share of its variance that is signal
  - The test of $\beta = 0$ stays valid; magnitudes are attenuated towards 0

Standardised Score: Correcting by the Correlation with the True Tone

- Standardised score z, non-differential error: $\mathrm{Cov}(r, z) = \beta\,\mathrm{Cov}(s^*, z)$
  - Non-differential error: $E[r \mid s^*, \hat{s}] = E[r \mid s^*]$
  - Since $\mathrm{Var}(z) = 1$, the slope is $b_z = \mathrm{Cov}(r, z) = \beta\,\sigma_{s^*}\,\rho(s^*, \hat{s})$
  - Slope on z = (effect of one standard deviation of s^*) $\times\ \rho(s^*, \hat{s})$: divide by the correlation of truth and score
  - A binary label is not classical error: misclassification shrinks a dummy slope by a different factor (Aigner (1973); Appendix and Seminar 15, A1)
- Econometrics of AI-generated variables
  - Battaglia et al. (2024): bias correction, joint estimation
  - Ludwig, Mullainathan & Rambachan (2024): prediction versus estimation uses of LLMs

How Much Attenuation in Our Slopes?

Score	$\rho(s^*, \hat{s})$, Twitter	b on day d (bp)	b/ρ (bp)	ℓ_j, one-factor loading
LM	0.29	15.5	54	0.53
FinBERT	0.58	39.2	67	0.85
Qwen2.5-7B	0.71	41.4	58	0.78

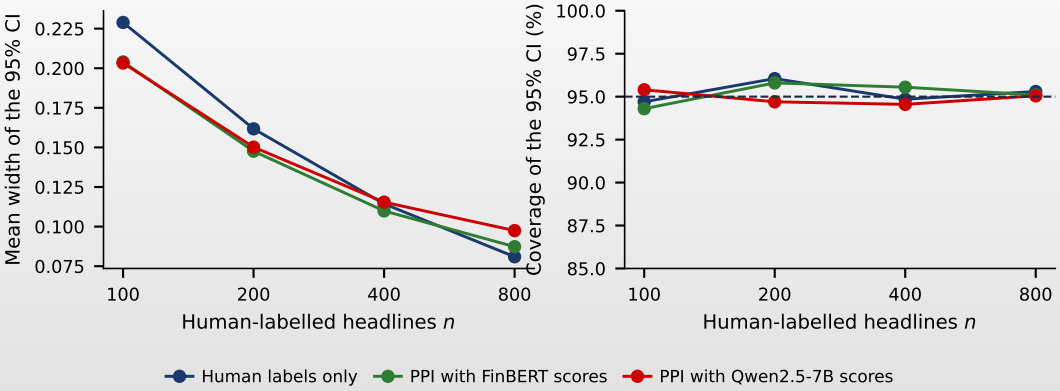
- ρ: correlation of the net score with the human label $(-1, 0, 1)$, one headline at a time
  - b/ρ: same-day effect of one standard deviation of true tone
  - After correction the three methods agree much more: the small LM slope is mostly measurement error, if the assumptions in Further Topics hold
- ℓ_j: loading of score j on the common factor f of the three daily scores
  - $\ell_j = \text{corr}(z_j, f)$ (factor model and instruments: Appendix)
  - Effect of one standard deviation of f, b/ℓ_j: 29 bp (LM), 46 bp (FinBERT), 53 bp (Qwen)

Further Topics: when correction overshoots

Valid Inference with Model Labels: Prediction-Powered Inference

- Data: n texts labelled by humans $(X_j,\ Y_j)$ and N unlabelled texts $\tilde{X}_i$
  - Independent random samples of one population
- Target: $\theta = E[Y]$, the population mean of the human label
  - $f(\cdot)$ is the score of a fixed model, trained on other data
- PPI (prediction-powered inference, Angelopoulos et al. (2023)): $\hat{\theta}_{\mathbf{PPI}} = \frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n}\sum_{j=1}^{n}\left(f(X_j) - Y_j\right)$
  - First term: the mean prediction over the N unlabelled texts
  - Second term: the mean bias $f - Y$, estimated on the labelled texts and subtracted
- $\widehat{\mathrm{Var}}(\hat{\theta}_{\mathbf{PPI}}) = \widehat{\mathrm{Var}}(f)/N + \widehat{\mathrm{Var}}(f - Y)/n$
  - An asymptotically valid CI under this sampling design, whatever the quality of f

PPI on Twitter Headlines: the Bias Correction Is the Main Gain (1/2)



MFM_ch15_valid_inference ▶ Interactive

PPI on Twitter Headlines: the Bias Correction Is the Main Gain (2/2)

- Target: mean human net tone $(-1, 0, 1)$
  - The 2 388 validation headlines act as the population
  - 2,000 replications of n labelled and 2 388 $-n$ unlabelled headlines drawn independently, with replacement
- True mean tone $\theta = 0.054$
  - Plain model means are biased: FinBERT -0.014 (wrong sign), Qwen2.5-7B $+0.101$
  - No sample size fixes this
- PPI keeps nominal coverage (94.3% at $n = 100$)
  - Mean width at $n = 100$: 0.229 with labels only, 0.204 with PPI (FinBERT)
  - Variance reduction $1 - \mathrm{Var}(f - Y)/\mathrm{Var}(Y)$: 24% (FinBERT), 27% (Qwen)
  - Modest because the models are noisy
- At $n = 800$ PPI is wider (0.087 vs 0.081)
  - With only 2 388 $-$ 800 unlabelled texts, $\mathrm{Var}(f)/N$ is no longer negligible
  - For the return regressions, the same labelled sample allows rectifying both OLS moments (Egami et al. (2023))
- Type of evidence: comparison of procedures (PPI, human labels only, the mean of the scores)

MFM_ch15_valid_inference

Research Question: Does an LLM Read News as the Market Does?

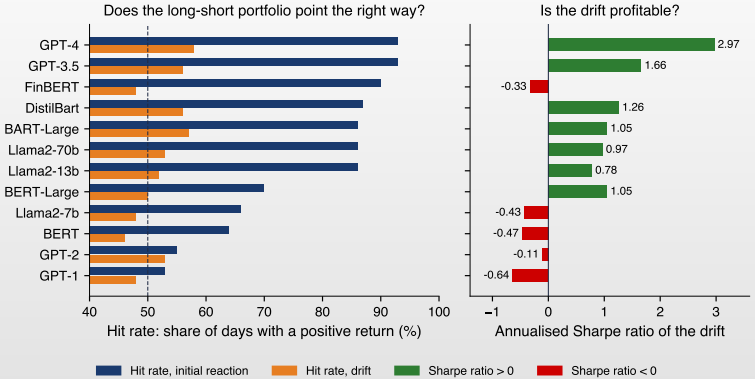
- Lopez-Lira & Tang (2026) (*Journal of Financial Economics*) ask:
  - Can a general-purpose LLM, with no explicit financial training, tell whether a headline is good or bad for the stock?
- Does its answer predict returns?
- Initial reaction: the price move right after the news
  - Not tradable: it measures whether the LLM understands the news
- Drift: the return after the reaction, in the direction of the news
  - Tradable: it measures underreaction of the market
- Versus our FNSPID test: the same zero-shot scoring of headlines
  - But on news after the model's training cutoff and with exact publication times
- Why it matters: the LLM becomes a measuring instrument for how fast prices absorb news
  - Their model: only LLMs above a quality threshold predict returns, and wide adoption should erode the predictability

GPT-4 Gets the Reaction Right and Earns the Drift

Model	Hit rate, reaction	Hit rate, drift	Drift, % per day	*SR*
GPT-4	0.93	0.58	0.34	2.97
GPT-3.5	0.93	0.56	0.29	1.66
FinBERT	0.90	0.48	−0.07	−0.33
GPT-1	0.53	0.48	−0.07	−0.64

- Overnight news, long-short portfolios: GPT-4 gets the direction of the reaction right on 93% of days
  - Like GPT-3.5 (version 3.5 of GPT), but it earns the largest drift
  - Short leg 0.26% per day (*SR* 2.01), long leg 0.08% (*SR* 0.78): the market underreacts mainly to bad news
- Regression of returns on the GPT-4 score, overnight news: $\hat{\gamma} = 1.326$ ($t = 26.52$) for the reaction, 0.161 ($t = 6.59$) for the drift
  - Adding the vendor's sentiment score leaves GPT-4 at 0.158 ($t = 5.27$) and the vendor at 0.010 ($t = 0.15$)
  - Small stocks: interaction coefficient 0.404 ($t = 4.55$): the drift is concentrated where arbitrage is costly
- Type of evidence: association
  - The GPT-4 Sharpe ratio falls from 6.54 (2021 Q4) to 1.22 (January–May 2024): Further Topics

Which Models Read the News Right, and Which Earn the Drift (1/2)



BART: Bidirectional and Auto-Regressive Transformer, an encoder–decoder language model; DistilBart: a compressed (distilled) BART; Llama2: open-weight LLMs released by Meta

MFM_ch15_llt_case ▶ Interactive

Which Models Read the News Right, and Which Earn the Drift (2/2)

- Long-short portfolios on overnight news, twelve models, same design
  - October 2021 – May 2024; source: Lopez-Lira & Tang (2026)
- Threshold: GPT-1 (Generative Pre-trained Transformer 1), GPT-2 (the first GPT versions), BERT and Llama2-7b get the reaction right on only 53–66% of days
  - They have negative drift Sharpe ratios (-0.64 to -0.11)
  - Seven models reach 86–93%, but only GPT-4 (2.97) and GPT-3.5 (1.66) pair it with a large drift Sharpe ratio; for every other model it is at most 1.26
- FinBERT gets the reaction right but has a negative drift
  - It gives a direction for only 25% of overnight headlines
  - Reading what is already in the opening price does not earn the drift
- Read the legs: stocks held on average in the short leg
  - 2 stocks for BERT-Large (Sharpe 1.05), 6 for GPT-3.5, 20 for GPT-4
  - A Sharpe ratio from a handful of shorts is fragile

MFM_ch15_llt_case

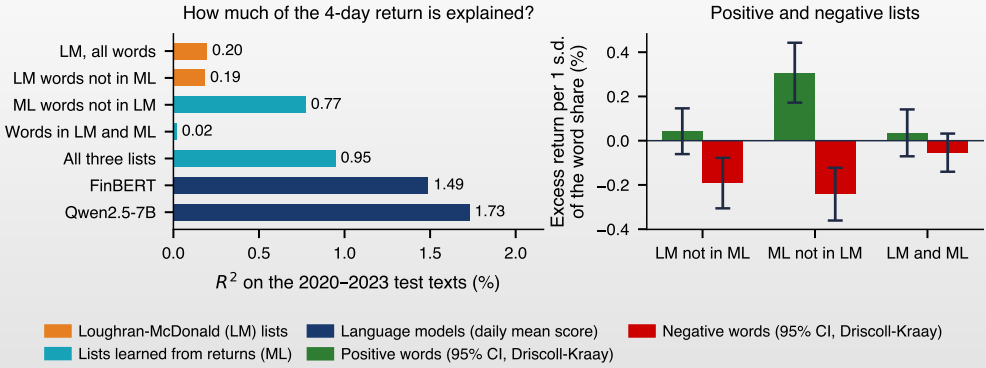
Replication in This Course, Limitations and an Open Extension

- The Qwen2.5-7B strategy of this chapter on 16 FNSPID stocks: +5.64 bp per day
  - Against 34 bp for GPT-4
  - Replication (Seminar 15, C3): the published prompt, temperature 0, scores $\{-1, 0, 1\}$, the two-firm rule (a leg only with at least two firms) and the same regression, on FNSPID headlines and the adjusted daily closes of the course data
  - Our data limits: date-only timestamps (the first safe return is $d + 2$, not the next open), 16 large stocks, headlines that predate the model
- Limits stated by the authors: at 20 bp per round trip the strategy loses money
  - So it suits low-cost traders such as market makers
  - Our 5 bp per one-way trade is 10 bp per round trip
  - Prompts and embedding models were not optimised, so the comparisons are lower bounds
- Open extension, on headlines published after the release of an open-weight model
  - does the drift survive?
  - does its Sharpe ratio keep falling as the authors' model predicts?

A Second Case Study: The Colour of Finance Words

- Garcia, Hu & Rohrer (2023) (*Journal of Financial Economics*), the question:
  - Can a word list chosen by price reactions compete with lists chosen by researchers, such as Loughran–McDonald (LM)?
- The paper's corpus: transcripts of 85 530 earnings calls (the calls in which firms present quarterly results), 3 229 firms, 13 Oct 2005 – 7 Oct 2020
- The label of each text: the stock's return minus the market return, from the close of day $t-1$ to the close of day $t+2$
- The paper's main result: lists learned on the calls before 2016, evaluated on 2016–2020, with the paper's control variables
  - Adjusted R^2: 2.1% with the LM words not in the ML (Machine Learning) lists, 4.6% with the ML words not in LM
  - With the controls only: 1.7%
  - The lists learned on the paper's full sample share only 18 positive and 30 negative words with LM
- Why it matters: the list stays transparent, but the colour of each word comes from the market, not from the researcher's judgement
  - It is the simple benchmark against which FinBERT and LLMs should be compared on the same texts

Word Lists Learned from Returns, on the FNSPID Headlines



ML (machine learning): the lists learned from returns; the lists are built on 2009–2019, R^2 and slopes are estimated on 2020–2023

MFM_ch15_colour_words ▶ Interactive

Method: Robust MNIR and What the Course Replication Changes

- MNIR (multinomial inverse regression, Taddy (2013), Appendix): the count of word j in text i is Poisson, with mean $m_i e^{\alpha_j + \varphi_j v_i}$
  - m_i: the number of words of the text; v_i: the standardised return
- The robust step: MNIR on 500 subsamples of 5 000 texts each, drawn without replacement
  - D_j^+: the share of subsamples with $\varphi_j > 0$ minus the share with $\varphi_j < 0$
  - Positive list: $D_j^+ \geq 80\%$; negative list: $D_j^+ \leq -80\%$
- The score of a text: the number of list words divided by the number of words, scaled to unit standard deviation
- The course replication, on the chapter's data:
  - The text: all FNSPID headlines of one stock on one day, 18 282 stock-days
  - Training: Oct 2009 – Dec 2019, without the texts whose return window reaches into 2020 (10 286 texts); test: 2020–2023 (7 980 texts)
  - Vocabulary (15 376 words) and winsorization cut-offs: from the training texts only
  - The sign of φ_j: a score-test approximation of the AIC selection (Akaike information criterion, $z^2 > 2$), instead of the gamma-lasso penalty
  - No control variables of the paper; the names of the 16 firms are removed instead of the proper-noun filter

The Result: Words Chosen by the Market Explain More than LM

- R^2 on the 2020–2023 test texts, slopes estimated on the same period: 0.20% with all LM words, 0.77% with the 94 ML words that are not in LM
  - The difference: +0.58 percentage points, 95% CI [0.14; 1.02] (bootstrap over blocks of 10 consecutive days, 1 999 draws, learned lists held fixed)
  - For cut-offs on D_j^+ from 50% to 95% (instead of 80%), the R^2 of the ML lists is between 0.44% and 1.02%
- One standard deviation more in the share of ML words (t with Driscoll–Kraay standard errors, HAC over days)
  - Positive: +0.31% ($t = 4.45$); negative: −0.24% ($t = -3.98$)
- The coefficient of the positive LM words is not significantly different from zero: +0.04% ($t = 0.81$)
  - Negative LM words: −0.19% ($t = -3.29$), significant also next to the ML lists: −0.17% ($t = -2.90$)
- FinBERT (1.49%) and Qwen2.5-7B (1.73%) explain more than any list
  - Qwen against the ML lists: +0.96 percentage points, 95% CI [0.20; 1.73]
- Type of evidence: a comparison of procedures, on an association within the same 4-day window, not a forecast
  - Many ML words describe the price move itself (“jumped”, “slumped”, “gainers”, “losers”): the headline may follow the return
- Limits: only 1 positive and 5 negative words shared with LM (positive; claims, downgrades, weak, worse, worst)
  - Qwen may already know what followed the 2020–2023 news (Look-Ahead Bias)

An Open Question: Can LLM-Generated Hypotheses Be Trusted?

- Ludwig & Mullainathan (2024): hypothesis generation as a statistical procedure
  - A model trained on data reveals a pattern, humans name it, fresh data test it
- Lopez-Lira, Tang & Zhu (2025): before its knowledge cutoff an LLM can recall outcomes
  - So a hypothesis it proposes about that period is not independent of the data that will test it
- Open: when an assistant proposes the signal, the prompt, the threshold and the holding period
  - what is the family of tests?
  - which inference stays valid?
- Agents that run the experiments themselves (Lu et al. (2024)) make the family larger and less visible

The Family of Tests: FWER and FDR

- Formally: $H_{\mathbf{0},j} : E[r_t^{(j)}] = 0$ for candidate strategies $j = 1, \ldots, M$
  - Control the FWER (Holm) or the FDR (Benjamini & Hochberg (1995))
  - FDR: expected share of false rejections among all rejections
  - BH (the Benjamini–Hochberg procedure) at level q: reject the hypotheses with $p_{(\mathbf{1})}, \ldots, p_{(k^*)}$, where $k^* = \max\{k : p_{(k)} \leq kq/M\}$
  - M counts every variant the assistant proposed, including those discarded; if the hypotheses were built on the test outcomes, the p-values need untouched test data or selective (sequential) inference

The Discovery Loop and What the Researcher Must Check

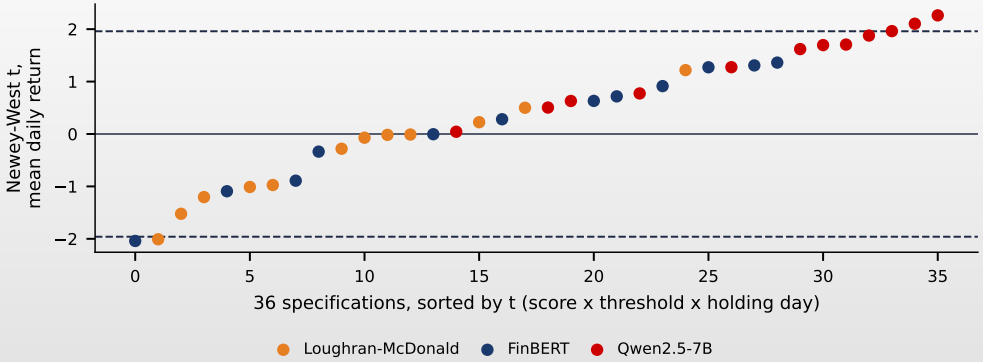
- Literature, hypotheses, data and code, robustness fixed in advance, critique, report
  - An assistant (Claude, ChatGPT, Gemini, GitHub Copilot) drafts each step
  - The researcher verifies each one (reusable prompts: Appendix)
- **Memorisation**: the model knows what followed the headline
  - Test: news after the release of the weights; anonymised text as a diagnostic (Glasserman & Lin (2023)); the memory test of this chapter
- **Forking paths**: prompt, threshold, horizon tuned on the test period
  - Test: a pre-registered grid and a stated error rate: Holm (FWER) or BH (FDR, under independence or positive dependence); $|t| > 3$ (Harvey, Liu & Zhu (2016)) is a heuristic, not an error-rate guarantee
- **HARKing** (hypothesising after the results are known)
  - A plausible mechanism written after the result, with references that may not exist
  - Test: state the mechanism and its falsifier before the test; resolve every DOI (digital object identifier)
- **Timing and agents that run code**
  - Date-only timestamps
  - Silent changes of data, windows or seeds
  - Test: first tradable return on $d + 2$; read-only data, logged tool calls, the diff of every script before a result is accepted

Mini-Case: Refinements Suggested by an Assistant

- A typical assistant proposal:
  - "the Qwen signal works; refine it with a stronger threshold or a later entry day"
  - We test the whole grid on the saved scores
- Grid: 3 scores (LM, FinBERT, Qwen) $\times$ 3 thresholds $\times$ 4 return days = 36 strategies
  - Thresholds $|s_{i,d}| > \delta$, $\delta \in \{0, 0.25, 0.5\}$, on the raw score
  - Return days $d+2$ to $d+5$, each a one-day position opened at the previous close
  - Base case: 5.64 bp per day, $t = 2.26$
- $|t| > 1.96$ for 5 of 36 strategies, against 1.8 expected under the global null at 5%
  - 2 of them have the wrong sign
- Survivors: Holm (FWER 5%) 0, BH (FDR 10%) 0, $|t| > 3$ 0
  - The Bonferroni threshold is $|t| > 3.20$
- Reading: the refinements add no evidence
  - The best variant is the one we started from
  - Itself chosen among three scores, on a period the model may remember

MFM_ch15_ai_discovery

The Specification Curve: Is It Different from Noise? (1/2)



MFM_ch15_valid_inference ▶ Interactive

The Specification Curve: Is It Different from Noise? (2/2)

- All 36 strategies of the mini-case sorted by t (Simonsohn, Simmons & Nelson (2020)); dashed: ± 1.96
- Joint null: no specification predicts returns
  - Flip the signs of the positions at random in blocks of 10 consecutive trading days, zero-return days included, 2 000 times, with the same flips for all 36 specifications (dependent wild bootstrap, Shao (2010))
  - Valid if, under the null, signed returns are symmetric around 0 and blocks are roughly independent
- Observed: 3 specifications with $t > 1.96$, median $t = 0.50$
  - Under the null, 3 or more occur with probability 0.113; a median this high with probability 0.207
- Reading: the curve is compatible with noise at 5%
  - Reporting only the best point would hide this
- Type of evidence: association
  - None of the 36 predictive associations stands out from noise

MFM_ch15_valid_inference

Your Turn: a Project Seed

- Question: does LLM headline sentiment predict returns beyond $d + 1$?
  - On news published after the model's release
  - Under a pre-registered grid
- Replicate first the Qwen strategy of this chapter, then the post-cutoff design of Lopez-Lira & Tang (2026)
- Deliverables
  - The pre-registered grid and correction; dated headlines after the release of an open model; the test on all variants
  - An LLM-proposed mechanism, stated before the test with its falsifier; a referee-style critique and your reply
- Log each prompt that shaped a decision in `AI_USE.md`
  - Log each invented reference, memorised fact or wrong number caught in `AI_ERRORS.md`, with the test that caught it (templates)

Key Takeaways

- Generic dictionaries misread finance
  - The LM dictionary is a transparent baseline
- FinBERT and 7B–14B open LLMs show no significant accuracy difference on financial headlines
  - Not proof of equality
  - A few hundred labels and a simple classifier compete with both
- LLMs remember market history
  - Evaluate them only after their release date
- Headline sentiment explains same-day moves
  - With realistic timing and costs it does not beat the market in the 16 large stocks studied
- Scores are measurements with error
  - Attenuation, possibly shared errors across models, few clusters and many tries all shape the inference

Check your understanding

- Why can a method that adds up the dictionary labels of isolated words misclassify a financial phrase?
  - Give an example in which context changes the sign.
- Can a high classification accuracy establish trading profitability?
  - Separate label quality, timing and transaction costs.
- Why must a historical news backtest consider the model's release date?
  - Describe the data choice that rules out the evaluated returns being memorised in the weights.
- Why is a sentiment score a measurement rather than an observed truth?
  - Explain what a random sample of human-labelled headlines allows you to estimate.

Short model answers: Appendix.

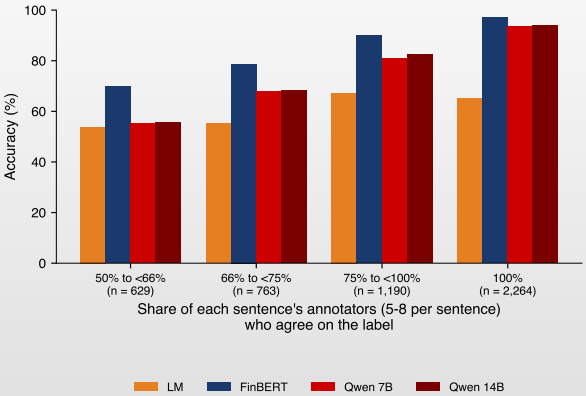
The Practical Decision

- **What I choose**
  - A fast, transparent baseline: the LM dictionary, no training and auditable
  - Short English financial headlines: FinBERT or, with a few hundred labels from the domain, embeddings + logistic regression
  - No labels, other languages or long documents: an open LLM of at least 7B, zero-shot
  - For daily trading of the 16 large stocks: no sentiment signal
- **Under which conditions**
  - Out-of-sample evaluation, with intervals and paired tests
  - On Twitter, TF-IDF + logistic regression reaches 83.2% accuracy, against 74.2% for Qwen2.5-7B and 72.5% for FinBERT
  - The Qwen2.5-7B strategy breaks even at 1.5 bp per trade, below the assumed cost of 5 bp
- **What result would make me drop it**
  - A significant McNemar test in favour of another method, on texts from the target domain
  - A memory test with an AUC above 0.5 over the backtest period (observed here: 0.69 for Qwen2.5-14B on the months before release)
  - A signal that stays significant after Holm on headlines published after the model's release, with a break-even cost above the real cost

Worked Example: One Sentence, Two Dictionaries

- Financial PhraseBank sentence, labelled **positive** by all annotators:
  - *Uponor improved its performance in spite of the decrease in residential construction in the US .*
- LM: positive words improved; negative words –
  - $\tau = (1-0)/(1+0) = +1.00$: positive, correct
- GI: positive words –; negative words spite, decrease
  - $\tau = (0-2)/(0+2) = -1.00$: negative, wrong
  - "Spite" and "decrease" describe the market, not the firm, and "improved" is not in GI

Annotator Agreement and Model Accuracy



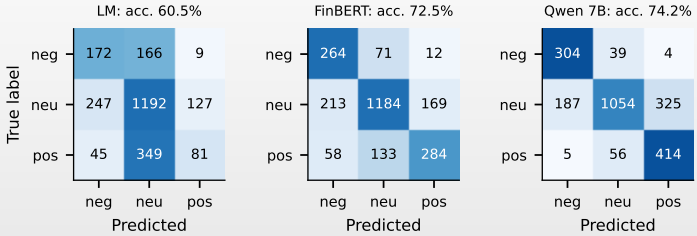
- Financial PhraseBank split by the highest agreement threshold each sentence reaches
  - Lowest group: 50% to below 66%
- Full agreement: FinBERT 97%, Qwen2.5-14B 94%, LM 65%
  - Agreement from 50% to below 66%: FinBERT 70%, Qwen2.5-14B 55%, LM 54%
- The human label is itself noisy: it caps measured accuracy
  - Latent-label models: Appendix
  - Symmetric noise, independent of the classifiers, shrinks measured differences towards 0; class-dependent noise can enlarge or reverse them
- For signals: a strong, unambiguous tone is more informative than a borderline one

MFM_ch15_classification

Agreement beyond Chance: Cohen's κ

- Accuracy also counts agreement reached by chance, which is large when one class dominates
- Cohen's κ (Cohen (1960)): $\kappa = \dfrac{p_o - p_e}{1 - p_e}$
  - p_o: the share of texts on which two raters (here the model and the human) give the same label
  - $p_e = \sum_k p_k^A \, p_k^B$: the agreement expected by chance, from the marginal label frequencies of the two raters
  - $\kappa = 1$: perfect agreement; $\kappa = 0$: no better than chance
- Twitter validation set: $\kappa = 0.19$ (LM), 0.50 (FinBERT), 0.56 (Qwen2.5-7B)
- The "always neutral" benchmark has $\kappa = 0$, although its accuracy beats that of LM
- For more raters per text: Fleiss' κ (Appendix)

Where the Errors Are: Different Models, Different Mistakes



- Twitter validation set: rows are the true labels, columns the predicted labels
  - Counts in the cells
- LM sends most positive headlines to neutral: its positive list is short
- FinBERT errs in both directions around neutral
- Qwen2.5-7B: recall of about 87% on positive (414/475) and 88% on negative headlines (304/347)
  - But precision of only 56% on positive (414/743) and 61% on negative (304/496)
  - Many neutral headlines are called positive or negative
- For trading, false positives cost money: precision on the extreme classes matters

MFM_ch15_classification

Same Model, Different Wording: the Prompt Is a Parameter



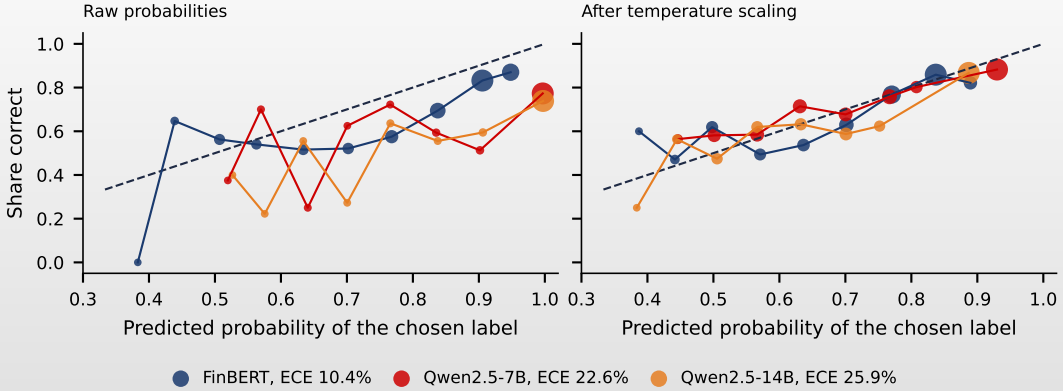
- Three prompts that ask the same question; Twitter validation set
- Accuracy ranges: 1.5B from 52.7% to 61.7%; 7B from 73.7% to 77.8%; 14B from 69.9% to 75.6%
  - The three prompts give the same label on only 67% of headlines (7B)
  - Share of "neutral" answers for 7B: 48% (P1) to 69% (P3)
- Choosing the prompt after seeing test results is a form of data snooping (Chapter 13)
  - Fix the prompt in advance, or choose it on a separate validation set

MFM_ch15_llm_zero_shot

Calibration of Probabilities: ECE and Temperature Scaling

- The question: when the model gives a class probability 0.9, is it right 90% of the time?
- ECE (expected calibration error) $= \sum_b \frac{n_b}{n} |\text{acc}_b - \text{conf}_b|$ over 10 bins of the top probability
  - n_b and n: the number of headlines in bin b and the total number of headlines
  - conf_b: the mean top probability in bin b
  - acc_b: the share of headlines classified correctly in bin b
- Temperature scaling (Guo et al. (2017)): $p_k(T) \propto p_k^{1/T}$, i.e. logits divided by T
  - T fitted by maximum likelihood on one random half of the Twitter validation set
  - ECE measured on the other half
  - The saved probabilities are rounded to 6 decimals; zeros are replaced numerically by 10^{-12}
  - The most likely class does not change, hence neither does accuracy

Are the Probabilities Calibrated? LLMs Are Overconfident (1/3)



MFM_ch15_valid_inference

Are the Probabilities Calibrated? LLMs Are Overconfident (2/3)

- Left panel: raw probabilities; right panel: after temperature scaling
  - The diagonal means perfect calibration
- ECE of the raw probabilities: FinBERT 10.4%, Qwen2.5-7B 22.6%, Qwen2.5-14B 25.9%
- Qwen2.5-7B gives 88% of headlines a confidence above 0.93 (mean 99.7%)
  - It is right on only 77% of them: overconfidence
- Temperature scaling, $T = 8.4$ (7B) and 1.4 (FinBERT): ECE falls to 4.3% and 4.6%
  - T for 7B depends on the numerical floor: 5.6 with 10^{-6} instead of 10^{-12}; the overconfidence remains
- ECE of 7B depends on the prompt: 22.6%, 22.1%, 19.8% under P1, P2, P3

MFM_ch15_valid_inference

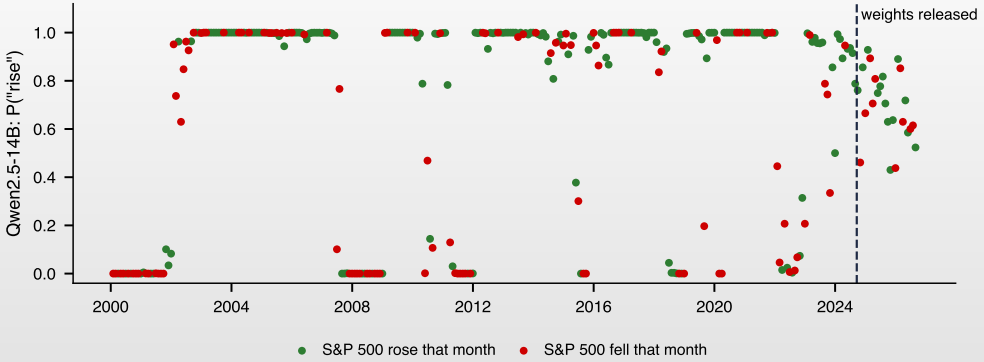
Are the Probabilities Calibrated? LLMs Are Overconfident (3/3)

- Two proper scores of the probabilities, lower is better
  - log loss: the mean of $-\ln p_{iy_i}$, with p_{iy_i} the probability given to the true class y_i of headline i
  - Brier score (Brier (1950)): squared distance between the probabilities and the one-hot label (1 for the true class, 0 otherwise), summed over the 3 classes
- For 7B, temperature scaling moves the log loss from 2.61 to 0.63 and the Brier score from 0.468 to 0.366
- With raw LLM probabilities the net score in a regression is close to a three-valued label

Worked Example: LLM versus FinBERT

- Twitter, prompt P1: Qwen2.5-7B 74.2%, FinBERT 72.5%
  - A = FinBERT, B = Qwen; texts that only one model gets right: $n_{01} = 395$, $n_{10} = 355$
  - Gap $(n_{01} - n_{10})/n = 40/2\,388 = +1.7$ points, CI 95% $[-0.5; +3.8]$; by hand, $\chi^2 = (|n_{01} - n_{10}| - 1)^2/(n_{01} + n_{10}) = 2.03$, so $p = P(\chi^2_1 > 2.03) \approx 0.15$ (exact test: 0.154)
  - Qwen2.5-14B against FinBERT: -0.1 points, $p = 0.942$
- No significant difference, which is not proof of equal accuracy
  - 70 to 130 times more parameters and slower processing bring no demonstrated gain
  - Where LLMs help: no labels needed, any language, longer and more complex documents (Appendix)

Memorisation Test for Qwen2.5-14B on the S&P 500 (1/2)



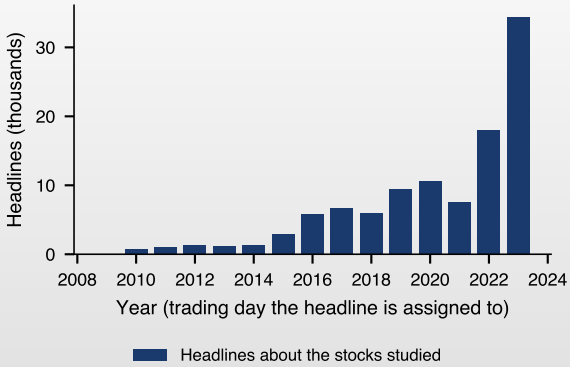
MFM_ch15_lookahead

Memorisation Test for Qwen2.5-14B on the S&P 500 (2/2)

- Probability of the answer "rise" for each month, no context given
  - Colour: what the index actually did
- "Fall" for 2000–2002 and 2008–2009, "rise" for most of 2003–2007 and 2013–2021
  - October 2008 (-16.9%), March 2020 (-12.5%), December 2018 (-9.2%): $P(\text{rise}) \approx 0$
- But September 2002 (-11.0%) and February 2009 (-11.0%) get $P(\text{rise})$ close to 1
  - Knowledge is coarse: memorable episodes rather than every month
- After the release the answers drift towards "rise"
  - No knowledge, only the general belief that markets go up

MFM_ch15_lookahead

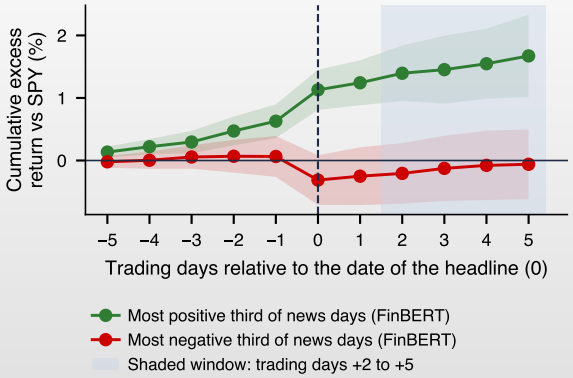
How Many Headlines, and When: Coverage Is Uneven



- Headlines about the 16 stocks per year, by the trading day they are assigned to
- 34 380 headlines in 2023 alone, against 8 445 for all years before 2016
  - Results are dominated by recent years; early years have few news days per stock
- On a news day a stock has on average 5.8 headlines
  - Daily score: the average of the headline scores for that stock and day
- Survivorship bias: the 16 stocks belong to today's large firms, selected with hindsight

MFM_ch15_news_data

Event Study: the Move Happens Before and On the News (1/2)



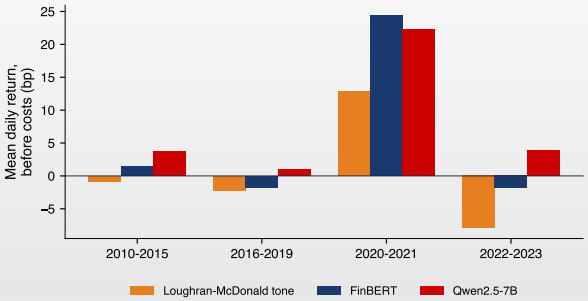
- Mean cumulative excess return, days -5 to $+5$
  - News days split by terciles of the daily FinBERT score
  - Shaded: tradable days $d+2$ to $d+5$
- Most positive third: $+0.63\%$ over the 5 days before, $+0.50\%$ on day d
  - Most negative third: $+0.06\%$ and -0.38%
  - Leaks, or headlines written after the price moved

MFM_ch15_news_signal

Event Study: the Move Happens Before and On the News (2/2)

- Sentiment effect = positive minus negative
  - Estimand: the difference in means, each stock-day event weighted equally; Driscoll–Kraay CI (HAC over days)
  - Day d: $+1.00\%$ $[+0.87; +1.13]$; days $d+2$ to $d+5$: $+0.17\%$ $[-0.11; +0.46]$
  - The chart averages within each day first, then over the days of each third: $+0.88\%$ on day d
- Test robust to event-induced variance (Boehmer, Musumeci & Poulsen (1991)), $d+2$ to $d+5$
  - $z = \overline{SCAR}/(s(SCAR)/\sqrt{N})$, with $SCAR_e = CAR_e/\hat{\sigma}_e$ the cumulative excess return of event e divided by its standard deviation estimated before the event; $N = 6\,094$ events per third
  - $z = +2.62$ (positive), $+1.25$ (negative) if events are independent; with Driscoll–Kraay standard errors: $+1.57$ and $+0.66$
  - Both thirds drift up; the selection of today's large firms may contribute

Stability over Time: Weak Evidence



- Mean daily return of the strategy before costs, by subperiod
- FinBERT: 2010–2015 +1.4 bp, 2016–2019 −1.8 bp, 2020–2021 +24.3 bp, 2022–2023 −1.8 bp
  - Only 2020–2021 is large, yet not significant at 5%: FinBERT $t = +1.89$, Qwen +22.2 bp ($t = +1.78$)
  - 2020–2021: the meme-stock period (GameStop episode: Appendix)
  - After the assumed costs the point estimates stay positive, but the statistical evidence is weak
  - Over the full sample the net result is negative
- Why so little: 16 heavily followed stocks, date-only timestamps, daily frequency
  - Published predictability relies on precise timestamps and many small stocks (case study)

MFM_ch15_news_signal

When Does the Correction Overshoot? Headlines versus Stock-Days

- The correction b/ρ rests on three assumptions
  - Non-differential error: given the true tone, the score error does not predict returns
  - The Twitter correlation carries over to FNSPID headlines
  - The same unit: ρ is measured per headline, but b uses the daily score of a stock, the mean of on average 5.8 headlines
- If headline errors are independent and the headlines of a stock-day share one true tone:
  - Averaging raises the correlation with the truth
  - The daily ρ exceeds the headline ρ, and b/ρ over-corrects
- Numerical example: one headline has reliability $\lambda = \rho^2 = 0.25$ ($\rho = 0.50$)
  - The mean of $k = 4$ such headlines has $\lambda_k = \dfrac{k\lambda}{1 + (k-1)\lambda} = \dfrac{4(0.25)}{1 + 3(0.25)} = 0.57$ (Spearman–Brown)
  - So $\rho_k = 0.76$
  - A daily slope of 10 bp is then corrected to $10/0.50 = 20$ bp with the headline ρ, but to $10/0.76 \approx 13$ bp with the daily one
- With independent headline errors our b/ρ values are upper bounds
  - The conclusion about the LM slope depends on this
- Remedy: labelling every headline of a random sample of stock-days measures ρ at the right level

When PPI Helps, and How It Extends to Regressions

- The PPI interval is narrower than labels alone when $\mathrm{Var}(f)/N + \mathrm{Var}(f - Y)/n < \mathrm{Var}(Y)/n$
  - Roughly $\mathrm{Var}(f - Y) < \mathrm{Var}(Y)$ when $N \gg n$: the model must explain part of the label's variation
- For a regression the rectifier applies to every moment condition that contains the predicted variable
  - Also when it is a regressor (Egami et al. (2023))
- Egami et al. (2023): design-based supervised learning
  - A doubly robust version that allows labelling probabilities to depend on the text
- The few hundred labelled documents are not only training data
  - They are the validation sample that makes downstream inference valid

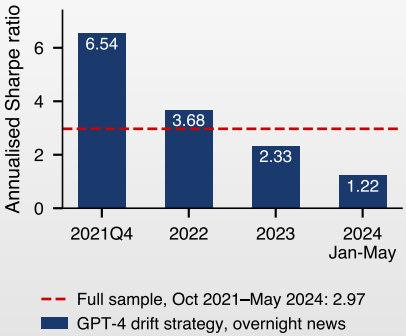
How the Authors Score News Published after the Training Cutoff of GPT-4

- 159,137 firm-headline-date observations, 4,123 US common stocks
  - Listed on the NYSE (New York Stock Exchange), NASDAQ (National Association of Securities Dealers Automated Quotations) and AMEX (American Stock Exchange) exchanges
  - October 2021 – May 2024
  - The sample starts after the September 2021 training cutoff of the scored model, `gpt-4-0314`
  - Filters: news-vendor relevance score 100, no "stock-gain"/"stock-loss" headlines, only new events, near-duplicates of the same day removed
  - Overnight news (before 9 a.m. or after 4 p.m.): 82% of headlines; the rest are intraday
- Prompt: the model acts as a financial expert
  - It answers `YES` (good news), `NO` (bad news) or `UNKNOWN` on the short-term effect on the stock price
  - Temperature 0
  - Score $s_{i,t}$: `YES` $= 1$, `UNKNOWN` $= 0$, `NO` $= -1$

Portfolios and Regression in Lopez-Lira and Tang

- Portfolios: each day, equal weights
  - Long the stocks with a positive score and short those with a negative one
  - A leg only with at least two firms
  - For overnight news, initial reaction: previous close to open; drift: open to close
  - Hit rate: share of days with a positive portfolio return; Sharpe ratio annualised from daily portfolio returns, as for our strategy
- Regression: $r_{i,t} = a_i + b_t + \gamma\, s_{i,t} + \varepsilon_{i,t}$
  - Firm and date fixed effects a_i, b_t
  - Standard errors clustered by date and by firm
  - γ: return per unit of score
  - Twelve language models of increasing size scored with the same design, from GPT-1 to GPT-4

The Drift Fades as LLMs Spread



- GPT-4 long-short drift strategy on overnight news, annualised Sharpe ratio by subperiod (source: Lopez-Lira & Tang (2026))
- From 6.54 in 2021Q4 to 1.22 in January–May 2024, more than a fivefold drop
  - The full-sample 2.97 pools a strong start with a weak end: not a forecast of today's performance
- Consistent with the erosion predicted by the authors' model, and the timing matches the spread of ChatGPT
- The authors also name changing market conditions
  - Short subperiods: the 95% intervals are wide, and the one for 2024 includes 0
  - Intraday news: the close-to-close drift is no longer significant in 2024

MFM_ch15_llt_case

When Annotators Disagree: Agreement and Latent Labels

- **Fleiss' κ** (Fleiss (1971)): N texts, m raters, n_{ij} raters giving text i label j
  - $\bar{P} = \frac{1}{N}\sum_i \sum_j \frac{n_{ij}(n_{ij}-1)}{m(m-1)}$, $\bar{P}_e = \sum_j p_j^2$, $\kappa = \dfrac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$; p_j: share of all votes for j
  - The formula assumes the same m for every text; PhraseBank has 5–8 raters per sentence and publishes only agreement levels, not the votes: κ cannot be computed from the release
- **Dawid–Skene model** (Dawid & Skene (1979)): rater a has a confusion matrix
  - $\pi_{jl}^{(a)} = P(\text{says } l \mid \text{true } j)$
  - The true class is latent
  - EM (expectation–maximisation) estimates the $\pi^{(a)}$ and gives each text a soft label $P(\text{true} = j \mid \text{votes})$
- Evaluate models by a proper scoring rule (log loss) against soft labels
  - Not by accuracy against the majority label
  - The ceiling in the agreement chart is a label-noise ceiling, not a model failure

Back

Worked Example: Is FinBERT Better than LM?

- Twitter validation set, 2 388 headlines: LM 60.5%, FinBERT 72.5%
  - LM wrong and FinBERT right: $n_{01} = 539$; LM right and FinBERT wrong: $n_{10} = 252$
  - Exact McNemar $p < 0.001$
- Difference +12.0 percentage points, bootstrap CI 95% [+9.8; +14.2]
  - Clear win for the fine-tuned model, even outside its training data

Back

Processing Time of the Qwen2.5 Models

- Wall-clock time per headline on a laptop graphics processor, including the prompt
  - Median of one pass per prompt
  - 0.5B: 4 ms; 1.5B: 12 ms; 3B: 20 ms; 7B: 92 ms; 14B: 105 ms
  - Single passes are noisy: 7B took 51 to 94 ms, 14B 88 to 112 ms; a timing claim needs repeated runs

Back

Beyond Headlines: Long Documents

- Annual reports: Loughran & McDonald (2011) built their lists from 10-K filings
  - Tone and uncertainty predict returns and volatility around filings
- Earnings calls and analyst reports: Huang, Wang & Yang (2023) show FinBERT extracts information that dictionaries miss
- Kim, Muhn & Nikolaev (2023): LLM summaries of disclosures are much shorter
  - But they explain stock-market reactions better than the originals
  - They measure disclosure "bloat": long, uninformative text is linked to lower price efficiency
- Long documents need models with long context windows, or chunking and aggregation

Back

Text Regressions with Statistical Theory

- Taddy (2013): multinomial inverse regression: model word counts given the outcome
  - $c_d \sim \text{MN}(q(y_d), m_d)$ (MN: the multinomial distribution)
  - c_d: counts, $m_d = \sum_j c_{dj}$: length
  - Log-linear model $q_j(y) \propto e^{\alpha_j + \varphi_j^\top y}$, loadings $\Phi = (\varphi_1, \ldots, \varphi_p)^\top$: given m_d and this model, $z_d = \Phi^\top c_d / m_d$ is sufficient for y_d
- Ke, Kelly & Xiu (2019) (SESTM, sentiment extraction via screening and topic modelling)
  - Screen words correlated with returns
  - Fit a two-topic model of positive and negative words
  - Score new articles by penalised likelihood
  - The paper gives statistical guarantees for the estimator; the signal is strongest on the day of and the day after the news
- Bybee et al. (2024): news attention to topics in 800,000 Wall Street Journal articles
  - It tracks economic activity and forecasts aggregate stock returns
  - Also Jegadeesh & Wu (2013) (word weights from 10-K reactions); rule for all: estimate on data before the forecast period

Back

Did the Memory Vanish? A Formal Test and Its Power

- H_0 : $\text{AUC}_{\text{before}} = \text{AUC}_{\text{after}}$, two independent samples
  - DeLong variance for each AUC (DeLong et al. (1988))
  - $z = (\widehat{\text{AUC}}_1 - \widehat{\text{AUC}}_2)/\sqrt{\hat{V}_1 + \hat{V}_2}$: 14B $z = 0.73$ ($p = 0.47$), 7B $z = 1.01$ ($p = 0.31$)
- Under H_0 (AUC $= 0.5$, no ties): $SE_0 = \sqrt{\dfrac{n_1 + n_0 + 1}{12\, n_1 n_0}}$
  - n_1, n_0: the numbers of up and down months
  - After release: $n_1 = 13$, $n_0 = 10$, so $SE_0 = \sqrt{24/1\,560} = 0.124$
- Smallest detectable drop (5% two-sided test, 80% power):
  - $2.80\sqrt{SE_{\text{before}}^2 + SE_0^2} = 2.80\sqrt{0.03227^2 + 0.124^2} = 0.36$
  - $2.80 = 1.96 + 0.84$, the normal quantiles for the test and for the power; SE_{before}: DeLong standard error of the 296 pre-release months
- A fall from 0.69 to 0.5 (0.19) is below this threshold
  - Not rejecting H_0 does not show that the memory survives release
- Months needed: the smallest m for which the detectable drop falls to 0.19
  - With the post-release share of up months kept at 13/23: $m = 101$ (7B: 129)

Back

Discrete Labels: Misclassification Is Not Classical Error

- Binary "negative news" dummy $s^* \in \{0, 1\}$, predicted label $\hat{s}$
  - $\alpha_0 = P(\hat{s} = 1 \mid s^* = 0)$, $\alpha_1 = P(\hat{s} = 0 \mid s^* = 1)$
  - The error $u = \hat{s} - s^*$ is negatively correlated with s^*: the classical formula does not apply (Aigner (1973))
  - plim $\hat{b} = \beta \dfrac{\pi(1-\pi)(1-\alpha_0-\alpha_1)}{p(1-p)}$, $\pi = P(s^* = 1)$, $p = P(\hat{s} = 1)$
- Twitter validation set, $\pi = 14.5\%$, from the confusion matrices
  - FinBERT: $\alpha_0 = 13.3\%$, $\alpha_1 = 23.9\%$, factor 0.45; Qwen2.5-7B: 0.59; LM ($\alpha_1 = 50.4\%$): 0.28
- As long as $\alpha_0 + \alpha_1 < 1$ the sign is kept, but a dummy slope can shrink to a third or less

Back

Two Noisy Measures: Factor Model and Instruments

- One factor: $z_j = \ell_j f + u_j$, $\mathrm{Var}(f) = 1$
  - u_j uncorrelated with f and with each other
  - Then $\rho_{jk} = \ell_j \ell_k$, so $\ell_j^2 = \rho_{jk}\rho_{jl}/\rho_{kl}$
  - Daily FNSPID scores: $\ell_j = 0.53$ (LM), 0.85 (FinBERT), 0.78 (Qwen)
- IV (instrumental variables): with $r = a + \beta f + \varepsilon$, instrument z_Q with z_F
  - $\hat{b}_{IV} = \widehat{\mathrm{Cov}}(r, z_F)/\widehat{\mathrm{Cov}}(z_Q, z_F) \to \beta/\ell_Q$
  - If $\ell_F \neq 0$, f is uncorrelated with ε, and u_F is uncorrelated with u_Q and ε
  - Day d, Qwen: 59 bp [53; 66] with FinBERT as instrument, 37 bp [28; 46] with LM (bootstrap over days)
- If both instruments were valid the two estimates would coincide
  - Difference 22 bp, $z = 5.4$: rejected
  - The test rejects the joint restrictions, not a named instrument: FinBERT–Qwen errors correlated through overlapping training text are one plausible cause, an LM error related to returns another; the factor loadings ℓ_j are then suspect too

Back

Social Media and Crowds: GameStop, 2021

- Bollen, Mao & Zeng (2011): collective Twitter mood predicted the Dow Jones in 2008
  - The result is sensitive to the sample and method
- January 2021: retail investors coordinated on social media
  - They pushed GameStop up by more than ten times in a few weeks
  - Attention, sentiment and trading rose together; this episode alone cannot separate prediction from causation: each needs its own test
- Lessons for text signals
  - Volume of attention can matter more than tone
  - Crowded signals revert and create risk for short sellers (Chapter 17: bubbles)



A GameStop store, Vallejo, California, February 2021

Photo: Will Buckner (2021); CC BY 2.0; Wikimedia Commons

Back

The Discovery Loop, Reusable Prompts (1/2)

- **Literature**
  - `"Find studies since 2023 that predict stock returns with LLM sentiment and test after the model's knowledge cutoff: DOI or arXiv id, cutoff date, test window, headline estimate."`
  - Resolve every identifier
  - Semantic Scholar, Google Scholar, Elicit for the search
  - Lopez-Lira & Tang (2026) is the reference design
- **Hypotheses**
  - `"Propose mechanisms by which headline sentiment could predict returns after d+1 and, for each, the cross-sectional pattern that would falsify it."`
  - Under-reaction predicts drift concentrated in small, less-covered stocks
  - Memorisation predicts an edge that disappears after the release date
- **Data and code**
  - `"From the daily news scores of this chapter and the daily closes of the course, reproduce the daily Qwen strategy of this chapter (enter at the close of d+1, hedge with SPY) and its Newey-West t."`
  - The scores are saved, so no model needs to be re-run

Back

The Discovery Loop, Reusable Prompts (2/2)

- **Robustness**, fixed in advance
  - `"Before any new run, write the full grid of variants you would test and the multiple-testing correction; all of them will be reported."`
  - Placebo: scores shuffled across stocks within the same day
  - Subperiods
  - Costs of 5 bp per trade
- **Critique**
  - `"Write a referee report: every way this strategy could reflect look-ahead, memorisation, selection or timestamps, with the test for each."`
  - Then check each proposed test yourself: a critique can also contain invented results
- **Report**: the whole grid, the variant chosen before the results, model versions and release dates

Back

AI Agents in Trading: Uses and Risks

- An agent is an LLM in a loop: read inputs, decide, call tools, observe, repeat
  - Tools: data, code, news search, order entry through an API, application programming interface
  - Used today for research summaries, coding and monitoring; fully autonomous trading is still rare and experimental
- Manipulation: a text read by the agent can carry instructions (prompt injection)
  - E.g. a fake press release
- Systemic risk (International Monetary Fund, IMF (2024)): many firms with similar models may trade alike
  - This raises the speed and correlation of moves in stress
  - Controls: position limits, human approval of orders, logging of every step, tests after each model update

Back

Checklist of Common Errors

- **Training overlap**: scored on its own training data, a model looks better than it is
  - FinBERT on PhraseBank
- **Look-ahead**: an LLM may remember the market after the news
  - Test after its release date
- **Timing**: date-only timestamps
  - Trade no earlier than the close after the news is surely public
- **Snooping**: prompts, thresholds and models chosen after seeing the test results
- **Costs and crowding**: small daily edges vanish after a few basis points of costs
- **Generated regressors**: scores are estimates
  - Correct attenuation
  - Validate on a random human-labelled sample (PPI)

Back

Short Answers to the Check Questions (1/2)

- **Dictionary**: a method that adds up isolated-word labels gives each word a single label
  - In "lower costs" it applies the negative label of "lower" without interpreting the phrase
  - In the usual context of firm profitability, lower costs can be favourable
  - "capital" or "board" are neutral business nouns that GI counts as negative
- **Accuracy and profit**: accuracy is measured against noisy human labels on other texts
  - A trade needs the first safe return ($d + 2$)
  - It must cover $4c$ a day and must survive the correction for many tries
  - Here all three scores explain same-day moves, yet no strategy earns a net profit after costs

Back

Short Answers to the Check Questions (2/2)

- **Release date**: before it, the evaluated returns may already be memorised in the weights
  - News and returns after the weight release could not have been memorised in them
  - Condition: no tool access to future information
  - The other sources of look-ahead bias in the backtest still need to be controlled
- **No-context memory test**: it can flag memorisation
  - Qwen2.5-14B: AUC 0.69 on the months before the release
  - A non-significant post-release AUC (0.59 on 23 months) does not establish the absence of contamination
- **Score as a measurement**: $\hat{s} = s^* + u$ estimates the true tone
  - Even a classical, zero-mean error can attenuate the slope (in the classical model, slope on z = effect $\times$ ρ)
  - A random human-labelled sample allows the mean bias $f - Y$ and the reliability of the score to be estimated
  - The estimates hold for the unit of analysis at which the sample was measured
  - PPI corrects the mean score using the mean difference $f - Y$
  - The b/ρ correction requires the model assumptions and a ρ measured at the same unit of aggregation as the regression
  - A ρ measured on headlines does not automatically justify correcting a slope on daily averages

Back

Text Analysis: Readings and Applications

- Tetlock (2007): media pessimism predicts market returns
- Loughran & McDonald (2011): a finance-specific sentiment dictionary
- Gentzkow, Kelly & Taddy (2019): text as data in economics
- Crypto applications
  - Crypto-asset classification Pele et al. (2023)
  - Dogecoin bubbles and sentiment Pele (2025)
  - Sentiment Quantlet Pele (2024)
- Further study: Quantinar: Natural Language Processing, Quantinar: DELTA, Quantinar: LDA Extensions
  - Textbook Franke, Härdle & Hafner (2019) for the statistics behind the tests

Back

Acronyms used in this material (1/2)

- **AI**: Artificial Intelligence
- **AIC**: Akaike Information Criterion
- **AMEX**: American Stock Exchange
- **API**: Application Programming Interface
- **AUC**: Area Under the (ROC) Curve
- **BART**: Bidirectional and Auto-Regressive Transformer
- **BERT**: Bidirectional Encoder Representations from Transformers
- **BH**: Benjamini–Hochberg (procedure)
- **BY-NC**: Creative Commons Attribution-NonCommercial
- **CI**: Confidence Interval
- **CLS**: CLaSsification token (BERT-type models)
- **DOI**: Digital Object Identifier
- **ECE**: Expected Calibration Error
- **EM**: Expectation-Maximisation (algorithm)
- **FDR**: False Discovery Rate
- **FNSPID**: Financial News and Stock Price Integration Dataset
- **FWER**: Family-Wise Error Rate
- **GI**: (Harvard) General Inquirer, dictionary IV-4
- **GPT**: Generative Pre-trained Transformer

Acronyms used in this material (2/2)

- **HARKing**: Hypothesising After the Results are Known
- **IMF**: International Monetary Fund
- **IV**: Instrumental Variables
- **LLM**: Large Language Model
- **LM**: Loughran-McDonald (financial sentiment dictionary)
- **LR**: Logistic Regression
- **MIT**: Massachusetts Institute of Technology
- **ML**: Machine Learning
- **MN**: MultiNomial (distribution)
- **MNIR**: MultiNomial Inverse Regression
- **NASDAQ**: National Association of Securities Dealers Automated Quotations
- **NYSE**: New York Stock Exchange
- **OLS**: Ordinary Least Squares
- **PCA**: Principal Component Analysis
- **PPI**: Prediction-Powered Inference
- **RLHF**: Reinforcement Learning from Human Feedback
- **ROC**: Receiver Operating Characteristic (curve)
- **SE**: Standard Error
- **SESTM**: Sentiment Extraction via Screening and Topic Modelling
- **TF-IDF**: Term Frequency - Inverse Document Frequency
- **US**: United States
- **UTC**: Coordinated Universal Time (Temps Universel Coordonné)
- **VADER**: Valence Aware Dictionary and sEntiment Reasoner

References (1/7)

- Aigner, D.J. (1973). Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics*, 1(1), 49–59.
- Angelopoulos, A.N., Bates, S., Fannjiang, C., Jordan, M.I., Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671), 669–674.
- Antweiler, W., Frank, M.Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance*, 59(3), 1259–1294.
- Araci, D. (2019). FinBERT: financial sentiment analysis with pre-trained language models. arXiv:1908.10063.
- Baker, M., Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4), 1645–1680.
- Battaglia, L., Christensen, T., Hansen, S., Sacher, S. (2024). Inference for regression with variables generated by AI or machine learning. arXiv:2402.15585.
- Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1), 289–300.
- Boehmer, E., Musumeci, J., Poulsen, A.B. (1991). Event-study methodology under conditions of event-induced variance. *Journal of Financial Economics*, 30(2), 253–272.
- Bollen, J., Mao, H., Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.

References (2/7)

- Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., et al. (2020). Language models are few-shot learners. *NeurIPS*; arXiv:2005.14165.
- Bybee, L., Kelly, B., Manela, A., Xiu, D. (2024). Business news and business cycles. *Journal of Finance*, 79(5), 3105–3147.
- Cameron, A.C., Gelbach, J.B., Miller, D.L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427.
- Cameron, A.C., Gelbach, J.B., Miller, D.L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2), 238–249.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Dawid, A.P., Skene, A.M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1), 20–28.
- DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3), 837–845.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: pre-training of deep bidirectional Transformers for language understanding. *NAACL*; arXiv:1810.04805.
- Dong, Z., Fan, X., Peng, Z. (2024). FNSPID: a comprehensive financial news dataset in time series. *Proceedings of the 30th ACM SIGKDD Conference*, 4918–4927.

References (3/7)

- Drinkall, F., Rahimikia, E., Pierrehumbert, J.B., Zohren, S. (2024). Time Machine GPT. arXiv:2404.18543.
- Driscoll, J.C., Kraay, A.C. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics*, 80(4), 549–560.
- Egami, N., Hinck, M., Stewart, B.M., Wei, H. (2023). Using imperfect surrogates for downstream inference: design-based supervised learning for social science applications of large language models. arXiv:2306.04746.
- Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Franke, J., Härdle, W.K., Hafner, C.M. (2019). Statistics of Financial Markets: An Introduction, 5th ed. Springer.
- Garcia, D., Hu, X., Rohrer, M. (2023). The colour of finance words. *Journal of Financial Economics*, 147(3), 525–549.
- Gentzkow, M., Kelly, B., Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574.
- Glasserman, P., Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. arXiv:2309.17322.
- Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q. (2017). On calibration of modern neural networks. *ICML*; arXiv:1706.04599.

References (4/7)

- Harvey, C.R., Liu, Y., Zhu, H. (2016). . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68.
- He, S., Lv, L., Manela, A., Wu, J. (2025). Chronologically consistent large language models. arXiv:2502.21206.
- Huang, A.H., Wang, H., Yang, Y. (2023). FinBERT: a large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2), 806–841.
- Hutto, C., Gilbert, E. (2014). VADER: a parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225.
- IMF (2024). Advances in artificial intelligence: implications for capital market activities. *Global Financial Stability Report*, October 2024, Chapter 3.
- Jegadeesh, N., Wu, D. (2013). Word power: a new approach for content analysis. *Journal of Financial Economics*, 110(3), 712–729.
- Ke, Z.T., Kelly, B., Xiu, D. (2019). Predicting returns with text data. NBER Working Paper 26186.
- Kearney, C., Liu, S. (2014). Textual sentiment in finance: a survey of methods and models. *International Review of Financial Analysis*, 33, 171–185.
- Kim, A., Muhn, M., Nikolaev, V. (2023). Bloated disclosures: can ChatGPT help investors process information? arXiv:2306.10224.

References (5/7)

- Lopez-Lira, A., Tang, Y. (2026). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Journal of Financial Economics*, 184, 104335.
- Lopez-Lira, A., Tang, Y., Zhu, M. (2025). The memorization problem: can we trust LLMs' economic forecasts? arXiv:2504.14765.
- Loughran, T., McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.
- Loughran, T., McDonald, B. (2016). Textual analysis in accounting and finance: a survey. *Journal of Accounting Research*, 54(4), 1187–1230.
- Lu, C., Lu, C., Lange, R.T., Foerster, J., Clune, J., Ha, D. (2024). The AI Scientist: towards fully automated open-ended scientific discovery. arXiv:2408.06292.
- Ludwig, J., Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *Quarterly Journal of Economics*, 139(2), 751–827.
- Ludwig, J., Mullainathan, S., Rambachan, A. (2024). Large language models: an applied econometric framework. arXiv:2412.07031.
- Malo, P., Sinha, A., Korhonen, P., Wallenius, J., Takala, P. (2014). Good debt or bad debt: detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4), 782–796.
- Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

References (6/7)

- Ouyang, L., Wu, J., Jiang, X., Almeida, D., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS*; arXiv:2203.02155.
- Pele, D.T. (2024). Sentiment Analysis (Quantlet, BitMood project). IDA Quantlets Repository, Bucharest University of Economic Studies.
- Pele, D.T. (2025). Dogecoin and speculative bubbles: a study using LPPL and sentiment analysis. Quantinar course, Bucharest University of Economic Studies.
- 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.
- Petersen, M.A. (2009). Estimating standard errors in finance panel data sets: comparing approaches. *Review of Financial Studies*, 22(1), 435–480.
- Qwen Team (2024). Qwen2.5 technical report. arXiv:2412.15115.
- Reimers, N., Gurevych, I. (2019). Sentence-BERT: sentence embeddings using Siamese BERT-networks. *EMNLP*; arXiv:1908.10084.
- Sarkar, S.K., Vafa, K. (2024). Lookahead bias in pretrained language models. SSRN Working Paper 4754678.
- Shannon, C.E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1), 50–64.

References (7/7)

- Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489), 218–235.
- Shiller, R.J. (2017). Narrative economics. *American Economic Review*, 107(4), 967–1004.
- Simonsohn, U., Simmons, J.P., Nelson, L.D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208–1214.
- Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503), 755–770.
- Tetlock, P.C. (2007). Giving content to investor sentiment: the role of media in the stock market. *Journal of Finance*, 62(3), 1139–1168.
- Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S. (2008). More than words: quantifying language to measure firms' fundamentals. *Journal of Finance*, 63(3), 1437–1467.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. (2017). Attention is all you need. *NeurIPS*; arXiv:1706.03762.
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M. (2020). MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained Transformers. *NeurIPS*; arXiv:2002.10957.
- Wu, S., Irsoy, O., Lu, S., Dabravolski, V., et al. (2023). BloombergGPT: a large language model for finance. arXiv:2303.17564.