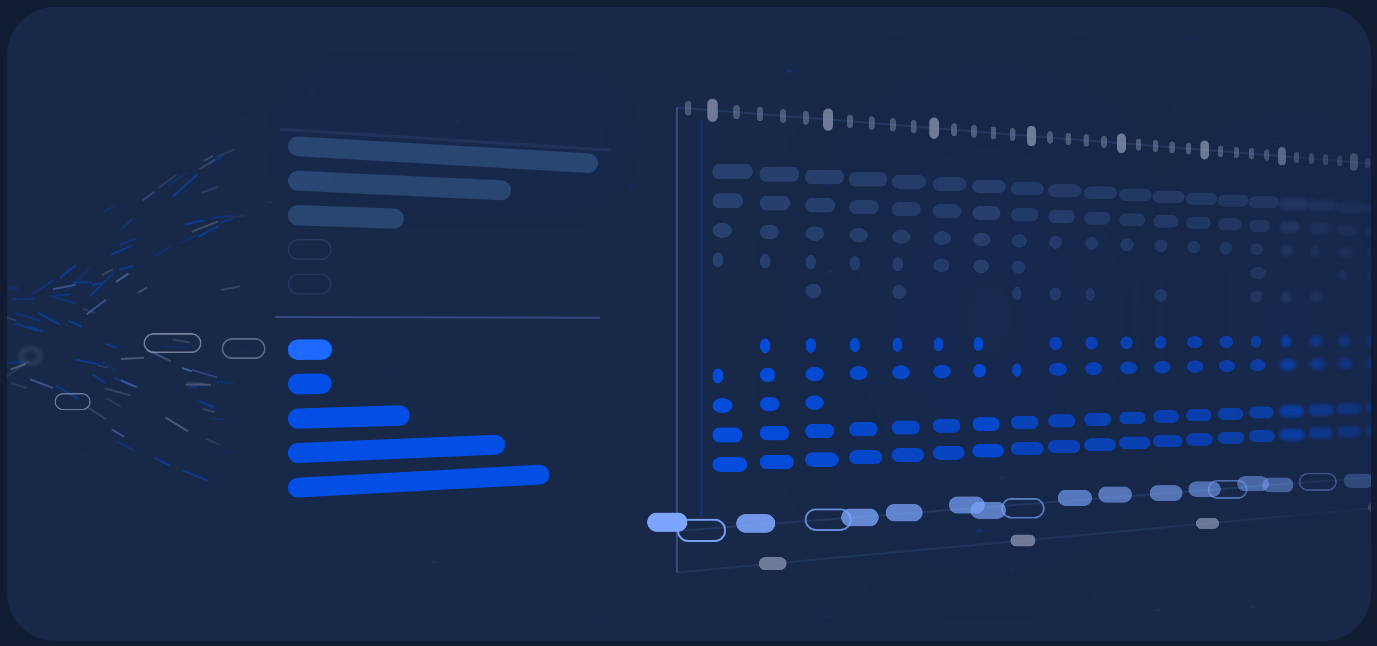
Rillor

RESEARCH REPORT

Deep learning on limit order books: what benchmarks show

Benchmark studies of deep learning on limit order books find lower scores on new stocks and periods than on FI-2010, and predictability that depends on a stock's tick size.



Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Benchmark and replication studies published from 2023 to 2025 find that the deep learning models for limit order books they retested score lower on new stocks and periods than on FI-2010, the public dataset their published scores came from, and that predictability depends on a stock's tick size [1, 2].

- FI-2010 holds ten days of limit order book (LOB) data for five stocks traded in Helsinki in June 2010, released normalized so the original book cannot be rebuilt [1, 3]. On it, the F1 classification score, which the papers report because the up, down and stationary labels are unbalanced, was about 46% for the dataset paper's baselines, and a 2025 transformer model reports an F1 of 92.81% at one horizon [1, 3, 4].
- LOBCAST, a 2023 benchmark study, retrained 15 deep learning models on the FI-2010 dataset and on Nasdaq data from 2021 and 2022 [1]. On the new data their F1 classification scores ranged from 48% to 61%, and the DeepLOB convolutional model scored 59.5% on both new datasets against a claimed 78.9% on FI-2010, averaged over horizons and seeds [1].
- In LOBCAST's 2023 retest of 15 order book models, about half of the hyperparameter search runs diverged to an F1 classification score of 33% or less, and the ranking of the models changed considerably once their claimed scores were set against the measured ones [1].
- A 2025 study that ran the DeepLOB model on 15 Nasdaq stocks from 2017 to 2019 tied predictability to tick size [2]. Its Matthews correlation coefficient (MCC), 0 for a random forecast and +1 for a perfect one, was 0.26 to 0.36 across three horizons for large-tick stocks, whose mean spread is at most about 1.5 ticks, and 0.01 to 0.11 for small-tick stocks, whose mean spread is about 3 ticks or more [2].
- In a 2025 study of 15 Nasdaq stocks, pT, the share of correct transactions among all those implied by the true labels and the forecasts, was 0.10 for CSCO at a 10-update horizon, where its MCC (a correlation score on which 0 means random) was 0.31 [2]. A 0.9 probability threshold raised BAC's MCC to 0.69 at a 50-update horizon while its pT fell to 0.02 [2].
- Order flow imbalance (OFI), the imbalance between supply and demand at the best bid and ask, explained 10-second changes in the mid-price, the average of the best bid and ask, in 50 U.S. stocks in April 2010 with an average R^2 of 65% in a linear model [5]. In a study of ten Nasdaq stocks from 2019 to January 2020, a deep model fed order flow was among the best models under a model confidence set test at 89% of the stock and horizon pairs where the authors found predictability, against 11% for the DeepLOB model fed order book states [6].
- LOB-Bench, a benchmark for generative models of order book messages presented at ICML 2025 [8], found in a downstream test on a subset of the models that adding their synthetic data did not improve a mid-price trend forecaster, and that one model's synthetic data reduced the forecaster's precision and recall [7].
- Benchmark papers from 2023 to 2025 recommend tests on other stocks and periods, published code and settings, labels tied to the tick or the spread, normalization that cannot leak future data and measures beyond the F1 classification score, and LOBCAST ran five random seeds per model and found the models sensitive to randomization [1, 2, 4].

Background

What a limit order book records

A limit order book records "all the outstanding bid and ask orders currently available on an exchange or a trading platform" [1]. The benchmark papers write the book's state as the prices and volumes of levels 1 to L on each side, a vector of 4L numbers [1, 2]. The mid-price is the average of the best bid and the best ask, and the spread is the best ask minus the best bid [1, 2]. On Nasdaq the tick, the smallest price step, "is fixed and equal to $0.01 for all the securities", with a qualification in a footnote [2].

Order book data comes at different depths. Lucchese, Pakkanen and Veraart separate L1, L2 and L3 data, with L2 holding "all available bid and ask prices and corresponding volumes" [6]. Nasdaq's TotalView-ITCH feed "uses a series of messages to track the life of a customer order", with timestamps in "nanoseconds since midnight" [9]. How a book is assembled from messages like these, and what one missed update does to it, is set out in rebuilding an order book from snapshots and diffs (https://rillor.com/insights/rebuilding-order-books-from-snapshots-and-diffs).

What the models are asked to predict

The FI-2010 line of models labels each moment up, down or stationary. The current mid-price m(t) is compared with the mean of the next k mid-prices: up if that mean is above m(t)(1 + θ), down if it is below m(t)(1 - θ), stationary otherwise [1]. The mean smooths the label, because comparing consecutive mid-prices directly "would result in a noisy labelled dataset" [1]. FI-2010 sets θ at "a threshold for the percentage change of 0.002" [3]. Briola, Bartolucci and Aste label by ticks instead: up or down if the mid-price moves at least one tick over H updates, stationary otherwise [2].

Time is counted in events. In Briola, Bartolucci and Aste's study "horizons are always defined in terms of LOB updates (which are unevenly spaced), while physical time is never used" [2]. For 13 of its 15 stocks, 10 updates "are more likely to happen in a physical time <1 s" [2].

Where the data comes from

Order book data is costly: "stock exchanges (e.g., NASDAQ) provide fine-grained data only for high fees" [1]. Two public sources recur in the equity studies reviewed here. Some use other data: DeepLOB's London Stock Exchange tests [12], the order flow imbalance paper's TAQ data [5] and Zhang, Lim and Zohren's market-by-order data [16]. FI-2010 is free and normalized (Finding 1). LOBSTER is the other: "The service LOBSTER provides to academic researchers is to reconstruct the historic order book from Nasdaq's Historical TotalView-ITCH data. For each selected ticker and date, LOBSTER returns message and order book files" [6]. The message file lists every market, limit and cancellation order with its time, type, id, size, price and direction [2]. LOBSTER data "are publicly available for the research community with an annual fee" [1], and Sirignano and Cont's data was "reconstructed from NASDAQ Level III data using the LOBSTER data engine" [10].

Early deep learning on order books

Sirignano's 2016 paper used a "spatial neural network" for the joint distribution of the book's future state, "trained and tested on nearly 500 stocks" [11]. DeepLOB, a convolutional model, followed in 2019 [12], and TransLOB, a 2020 preprint, claimed "a new state-of-the-art benchmark for the FI-2010 dataset" [13]. The same year, Briola, Turiel and Aste set simple networks against complex ones (Finding 3) [14].

Method

What was examined. Published evidence on how deep learning models for limit order books perform when tested away from the data their scores were first reported on, and how the later benchmark papers say such models should be evaluated.

Sources. The FI-2010 paper and its data record [3, 15]. Model papers from 2016 to 2026 [4, 10, 11, 12, 13, 16, 17]. Benchmark and replication studies [1, 2, 6, 14, 18, 19]. The order flow imbalance paper [5] and the abstract of a deep order flow study [20]. Generative models and simulators [7, 8, 21, 22, 23, 24, 25, 26, 27]. Code repositories [28, 29, 30, 31] and Nasdaq's TotalView-ITCH specification [9]. Thirty-one sources, each opened on 8 November 2026. The publisher's page for Kolm, Turiel and Westray's paper refuses automated access, so its abstract was read from its Crossref record, https://api.crossref.org/works/10.1111/mafi.12413 [20]. Several papers are preprints, among them TLOB, TensorLOB and LOB-Impact [4, 17, 27].

How. Figures come from the table or passage named beside them. Horizons stay in each paper's own units: FI-2010 labels the next 1, 2, 3, 5 and 10 of its 10-event blocks [1, 3], DeepLOB reports k = 10, 20, 50 and 100 [12], and Briola, Bartolucci and Aste use H in {10, 50, 100} updates [2]. No horizon is equated across papers, no score is set against another paper's dataset, label or metric, and no new averages were computed. No experiments were run for this report.

Findings

1. The FI-2010 benchmark and its limits

FI-2010 holds "five stocks from the NASDAQ Nordic stock market for a time period of ten consecutive days, leading to a dataset of ≈4,000,000 time series samples in total" [3]. The stocks, Kesko, Outokumpu, Sampo, Rautaruukki and Wärtsilä, traded in Helsinki from 1 June to 14 June 2010, with events kept between 10:30 and 18:00 [3]. Sampled every 10 events, the set holds "a total of 394,337 events" [1].

The data was normalized before release: "The open-access version of our data has been normalized in order to prevent reconstruction of the original NASDAQ data" [3]. The record lists z-score, min-max and decimal-precision versions, nine cross-validation folds and labels for five classification problems, under the Creative Commons Attribution 4.0 licence [15].

The dataset paper's two baselines reached "an average out-of-sample performance (F1) of, approximately, 46%" [3]. F1 is the focus because the classes are unbalanced [1, 3], and the balance shifts with the horizon:

Horizon	Up	Stationary	Down
k = 1	18%	63%	19%
k = 5	32%	35%	33%
k = 10	37%	25%	38%

Source: [1], Section 4.1.

θ was set to balance the classes at k = 5, so it "is not chosen in accordance with its financial implication" [1]. LOBCAST followed the split of the 13 published models it retested, the first seven days for training and validation and the last three for testing [1], which DeepLOB's paper calls Setup 2 [12].

Three later papers question the dataset. LOBCAST: it is "already pre-processed (filtered, normalized, and labelled) so that the original LOB cannot be backtracked" [1]. Briola, Bartolucci and Aste: it "is too simplistic, leaving ample space for models' overfitting", and using it as a benchmark "represents the first cause of the so-called 'simulation-to-reality' gap" [2]. DeepLOB's own authors: "it is far too short, downsampled and taken from a less liquid market" [12].

Scores on it kept rising. LOBCAST notes that recent models "have achieved over 88% in F1-score" [1]. TLOB, a 2025 transformer, reports F1 on FI-2010 of 81.55 at h = 10, 82.68 at h = 20, 90.03 at h = 50 and 92.81 at h = 100 (Table 9) [4]. The dataset is still in use: TensorLOB, posted on 2 November 2026, reports results on FI-2010 only, with a weighted F1 of 79.1 at h = 100 (Table 1) [17].

2. DeepLOB and the universal-feature claims

DeepLOB, by Zhang, Zohren and Roberts, is a deep convolutional neural network whose abstract says it "outperforms all existing state-of-the-art algorithms on the benchmark LOB dataset" [12]. Its reported F1 scores:

Test data	k = 10	k = 20	k = 50	k = 100
FI-2010, Setup 2 (Table II)	83.40%	72.82%	80.35%	
FI-2010, Setup 1 (Table I)	77.66%		74.96%	76.58%
London, five training stocks (Table IV)		70.15%	63.49%	60.65%
London, five stocks not in training (Table IV)		68.48%	62.84%	60.77%

Source: [12]. Blank cells are horizons a table does not report.

The London data covered five stocks on "all trading days from 3rd January 2017 to 24th December 2017", "more than 134 million samples" [12]. From the near-equal scores on held-out stocks the authors concluded that the model can "extract universal features" [12]. Its trading test assumed "trading takes place at mid-price" and compared "gross profits before fees" [12].

Sirignano and Cont tested the idea at scale, on "approximately 1000 stocks traded on the NASDAQ between January 1, 2014 and March 31, 2017", predicting "whether the next price move is up or down" [10]. A coin flip scores 50%, and "even a small deviation (i.e. 1%) from this 50% benchmark is statistically significant" [10]. Out of sample, in a comparison on about 500 stocks for June to August 2015 (Figure 4), stock-specific deep models gained "between 5% to 10% for most stocks" in accuracy over stock-specific linear models, and the universal model improved accuracy "by around 10%" over the linear models [10]. The results "also hold for stocks which are not part of the training sample", and splitting training data by "large/small tick stocks" did not improve results [10].

The two lines of work measure different things: two-class accuracy on the next move [10], against three-class F1 [1, 12]. Their numbers cannot be set side by side.

3. What replication found: LOBCAST

LOBCAST, by Prata and co-authors, retrained "fifteen state-of-the-art DL models": 13 published "between 2017 and 2022" and 2 baselines, all "originally trained and tested on the FI-2010 dataset", each with five seeds per horizon [1]. It added two LOBSTER datasets of six Nasdaq stocks, LOB-2021 (1 to 15 July 2021) and LOB-2022 (1 to 15 February 2022), each of 10 trading days, built "closely" to FI-2010's process [1].

Model	Claimed on FI-2010	Measured on FI-2010	LOB-2021	LOB-2022
DeepLOB	78.9 ± 4.4	71.4 ± 5.3	59.5 ± 3.0	59.5 ± 2.9
TransLOB	87.3 ± 4.0	59.4 ± 2.6	57.7 ± 2.9	50.4 ± 6.1
BiN-CTABL	80.1 ± 6.9	82.6 ± 7.0	61.2 ± 2.7	59.2 ± 3.3

Source: [1], Table 2, F1 in %, mean and standard deviation over horizons and seeds.

On the new data, performance was "significantly lower than on the FI-2010 dataset, ranging 48-61% in F1-score", and BiN-CTABL, which matched its claim on FI-2010 with a robustness score of 99.7, lost "approximately 19.6% in F1-Score across all horizons" [1]. ATNBoF was claimed at 67.1 ± 5.5 and measured at 40.9 ± 7.7 on FI-2010, and the MLP baseline at 51.8 ± 3.2, measured at 48.0 ± 2.6 on FI-2010 and 55.5 ± 3.9 on LOB-2021 [1].

The results were fragile. "All models are very sensitive to hyperparameters, in fact, they diverged (F1-score ⩽ 33%) during the hyperparameters search for about half of the runs" [1]. "The ranking of systems changes considerably" against the declared scores, though "the best six systems in FI-2010 remain the same in LOB-2021 and 2022" [1]. Ensembles "do not exceed the performance of the top-performing models", and five models showed "a standard deviation over the runs that exceeds 5 points" [1]. The authors give two possible reasons for the gaps: "the lack of the original code and the missing hyperparameters declaration" [1]. The models were "slightly biased toward the stationary class", and a trading simulation on LOB-2021 indicated that "the models' profitability is far from guaranteed" [1]. The top three models skipped the long inputs many papers use: "none of the top three models in our study utilize h = 100 long market observations as input, despite it being a common practice" [1]. Training took "approximately 155 hours for FI-2010 and 258 hours for LOB-2021/2022" on 8 GPUs, and the framework is public [1, 29].

The authors conclude that the models "are not consistently reliable as they often exhibit non-robust and non-generalizable performance" [1]. They name their limits: the grid search "is not exhaustive", and they could not train on datasets spanning years [1].

Other studies found simple models holding their own. On Intel LOBSTER data from 2019, "the Multilayer Perceptron performs comparably to or better than state-of-the-art CNN-LSTM architectures" [14]. On bitcoin order books, "simpler models can match and even exceed the performance of more complex networks" after preprocessing and tuning [19]. TLOB's authors report that a "simple MLP-based architecture" surpassed earlier results [4].

4. Tick size and predictability: LOBFrame

Briola, Bartolucci and Aste ran DeepLOB on LOBSTER data for 15 Nasdaq stocks from 2017 to 2019, with "45 consecutive days of training, 5 days of validation and 10 consecutive days of testing" each year and a "5-days feature-wise rolling window z-score normalization" [2]. The study ran 135 experiments with a "cumulative GPU runtime of 959 hours" [2]. Its code is public as LOBFrame [18, 30].

Stocks were grouped by mean spread: small-tick at about 3 ticks or more (CHTR, GOOG, GS, IBM, MCD, NVDA), large-tick at about 1.5 ticks or less (BAC, CSCO, KO, ORCL, PFE, VZ), and medium-tick in between (AAPL, ABBV, PM) [2]. MCC "takes values between -1 (in case of inverse prediction) and +1 (in case of perfect prediction), while a value of 0 indicates a random prediction" [2]. Without a probability threshold:

Tick class	H = 10	H = 50	H = 100
Small-tick	0.11	0.04	0.01
Medium-tick	0.13	0.085	0.036
Large-tick	0.29	0.36	0.26

Source: [2], Section 7.1. H counts order book updates.

For small-tick stocks at H = 100, "These results suggest that the model is producing random forecasts" [2]. At H = 10 for small-tick stocks, "the 29% of true class 1 is misclassified as class -1, and the 27% of true class -1 is misclassified as class 1" [2]. Large-tick stocks reached F1 above 0.45 without thresholds and above 0.7 with them, and accuracy above 0.7 without thresholds and over 0.9 with them [2].

The authors find "a direct mapping between the IR score of a stock and its tick size", so "the tick size itself could be used as a proxy measure of a stock's predictability" [2]. The information richness score comes from Kolm, Turiel and Westray [20]. This sits beside Sirignano and Cont's finding that splitting training data by tick did not help [10]. The two answer different questions: how to pool training data, and how predictable each class of stock is [2, 10].

5. From classification scores to complete transactions

Briola, Bartolucci and Aste argue that backtesting model outputs "by using historical data-only is not possible", because it assumes forecasts acted on in time, every order executed, and zero market impact and transaction costs [2]. They score forecasts by the transactions they imply. Potential transactions PT are counted on the true labels, executed transactions TT on the predictions and correct transactions CT where both agree, and pT = CT / (PT + TT - CT) [2]. As they put it, "we are more interested in the chronological location of prediction errors rather than in the number of their occurrence" [2].

Horizon	Stock	Tick class	pT	MCC	F1
H = 10	CSCO	Large	0.10	0.31	0.50
H = 10	GOOG	Small	0.04	0.20	0.46
H = 10	AAPL	Medium	0.15	0.17	0.42
H = 50	CSCO	Large	0.17	0.36	0.57
H = 50	BAC	Large	0.11	0.43	0.61
H = 50	NVDA	Small	0.01	0.01	0.19

Source: [2], Tables 8 and 9, no probability threshold, on the first 50% of the available data.

At H = 100, BAC reached a pT of 0.19 with an MCC of 0.35, and GOOG a pT of 0.06 with an MCC of 0.02 (Table 10) [2].

Thresholds move the measures apart. "pT decreases for increasing probability thresholds and increases moving from small-tick stocks to large-tick stocks", while "the MCC increases for increasing probability thresholds" [2]. For small-tick stocks "the probability of correctly executing a trade at a threshold larger than 0.5 is zero", net of minor oscillations [2]. At H = 50 with a 0.9 threshold, BAC's MCC was 0.69 and its pT 0.02 (Table 9) [2]. The preprint version states that "traditional machine learning metrics fail to adequately assess the quality of forecasts in the Limit Order Book context" [18].

Labels tied to trading costs show the same gap. With θ set to the average spread on Tesla, TLOB's F1 at h = 50 fell from 43.48 (Table 4) to 41.39 and was 30.82 at h = 200 (Table 8), and on Intel the method "would result in 99.99% of trends classified as stationary" [4]. A wide stationary band can also yield "a trivial 'always predict stationary' model with 95% accuracy" [19].

6. Order flow imbalance as the simple baseline

Cont, Kukanov and Stoikov found that "over short time intervals, price changes are mainly driven by the order flow imbalance", with "a linear relation between order flow imbalance and price changes, with a slope inversely proportional to the market depth" [5]. They used "one calendar month (April, 2010) of trades and quotes (TAQ) data for 50 stocks" from the S&P 500 [5], the kind of record described in TAQ data explained (https://rillor.com/insights/taq-data-explained). Across 273 half-hour subsamples per stock, the linear fit for 10-

second mid-price changes had an "average R2 of 65%" (Table 2), and a quadratic term moved R^2 "from 65% to 68% on average" [5]. The relation is contemporaneous: it explains price changes in the same interval and makes no forecast [5].

Order flow inputs did well in later studies. Kolm, Turiel and Westray's abstract, on 115 Nasdaq stocks, reports that "models trained on order flow significantly outperform most models trained directly on order books" and that "the effective horizon of stock specific forecasts is approximately two average price changes" [20]. Lucchese, Pakkanen and Veraart took ten stocks from a universe of "115 Nasdaq tickers from January 2, 2019, to January 31, 2020" and compared models with model confidence sets, "a formalized statistical inference framework" [6]. They count predictability for a stock at a horizon when at least one order book model statistically outperforms an unpredictive benchmark (Definition 4.1) [6]. The share of the stock and horizon pairs with identified predictability at which each model was in the superior set, at a significance level of 0.01 (Table 5) [6]:

Model and input	Share of stock and horizon pairs
deepOF, L2 order flow	89%
deepVOL, L3 volumes	86%
deepVOL, L2 volumes	84%
deepLOB, L2 book states	11%
Benchmark	0%

Source: [6], Table 5.

The authors add that "L3 data might be excessively granular when predicting high-frequency returns from order books" [6]. On market-by-order data, Zhang, Lim and Zohren found that order-level and book-level models "individually provide similar performance", and ensembles of the two can improve on them [16].

7. Generative models and market simulators

Replaying history "does not capture the market response to the experimental agents' actions" [24], and LOBCAST calls simulation "beyond testing on historical data" essential [1]. The queue-reactive model treats "the limit order book as a Markov queuing system", for use "as a market simulator or as a tool for the transaction cost analysis" [22]. ABIDES simulates "tens of thousands of trading agents interacting with an exchange agent", with messages "modeled after NASDAQ's published equity trading protocols ITCH and OUCH" [23]. A conditional generative adversarial network trained on aggregate historical data was set inside

ABIDES as a world agent [24], and an autoregressive model generates tokenized order book messages, trained on LOBSTER data of Nasdaq equity books [21]. Studies of LLM agents in simulated markets are reviewed in LLM agents as market participants (https://rillor.com/insights/llm-agents-as-market-participants), and measuring LLM agents in simulated auctions (https://rillor.com/insights/measuring-ai-agents-in-simulated-auctions) sets out how to test them.

LOB-Bench, presented at ICML 2025, evaluates "the quality and realism of generative message-by-order data for limit order books (LOB) in the LOBSTER format" [7, 8]. It measures distributions of spread, volumes, order imbalance and inter-arrival times, "scores from a trained discriminator network" and "market impact metrics" [7]. It tests four generative models and "a widely-used classic model as a baseline (Cont et al., 2010)" on GOOG and INTC [7]. Its library is public [31].

"The LOBS5 model beats the baseline and coletta model on almost all scores" (Figure 3), and the baseline "performs well on LOB depth and level-related scores, and much worse on time and volume metrics" (Figure 6) [7]. Errors grow as "the distance scores increase for longer unrolls", and a discriminator separated the baseline's generated data from real data with a ROC score of around 1 [7]. In a downstream test, "for the Coletta model in particular, including generated synthetic data markedly reduces the resulting prediction precision and recall" (Figure 8, a subset of the generated models) [7]. The other models showed the same pattern "particularly for short prediction horizons, although to a less extreme extent", and mixing in data from the baseline or LOBS5 "has no significant effect for longer prediction horizons" [7]. The authors concluded in 2025 that "current generative models are not yet able to generate data aiding this mid-price trend prediction task" (Section 6) [7].

Later work builds on it. A September 2025 diffusion preprint reports "state-of-the-art performance on LOB-Bench", by its own account [26]. TRADES (ECAI 2025) scores generators by a predictive score "measured as an MAE" [25]. LOB-Impact, posted on 19 October 2026, found impact in five generative models "with exponents in the range of 0.63 to 1.21, slightly above the traditional 0.5 of the square-root law of impact", and did not observe impact decay during relaxation [27].

8. What evaluation the benchmarks recommend

Practice	What the papers say
Test on unseen stocks and periods	Generalizability "on unseen market scenarios using LOBSTER data" [1], and "a cross-exchange validation" [2]
Characterize the stocks first	Estimate "a given stock's predictability based on its LOB microstructural properties" [2]
Publish code and settings	"publish the code developed" [1]. Code for DeepLOB, LOBCAST, LOBFrame and LOB-Bench is public [28, 29, 30, 31]

Practice	What the papers say
Run several seeds	LOBCAST ran five seeds per model and horizon and found the models "susceptible to hyperparameter selection, randomization" [1]
Tie labels to costs	"a dynamically adapting θ" [1], the average spread [4] or one tick [2]
Normalize without leakage	A rolling 5-day z-score "prevents any data leakage" [2]
Score beyond F1	MCC and pT [2], and model confidence sets [6]
Test generative models three ways	Distributions, impact and downstream usefulness [7, 25, 27]

Scores also decay with time. TLOB's F1 at h = 50 on Intel was 66.87 on one 2012 day and 60.19 in January 2015, though the 2012 test was a single day because "it was the only day available" (Table 7) [4]. Speed matters too: LOBCAST measured inference at 1.31 ms for DeepLOB and 0.08 ms for the MLP (Table 1) [1].

Licensing limits replication. LOBCAST could not release its datasets "Due to copyright reasons" [1], TLOB's authors "cannot make the dataset public for copyright reasons" [4], and FI-2010 was normalized to prevent reconstruction [3].

Implications

For researchers publishing a new model. A score on FI-2010 alone describes ten days of five Helsinki stocks from 2010, released already filtered, normalized and labelled [1, 3]. The later papers ask for other stocks and periods, code and hyperparameters and a label tied to the tick or the spread, and LOBCAST reports each model over five seeds [1, 2, 4]. A result is easier to judge beside an MLP and an order flow input under the same split, since both matched or beat complex models in several studies [6, 14, 19].

For quant teams reading published results. Check the horizon unit, the label, the split and whether a score is claimed or reproduced [1, 2, 3, 12]. Ask which stocks were tested and their tick sizes, since MCC for one model ran from 0.01 to 0.36 by tick class and horizon [2]. Read F1 beside a transaction measure such as pT, because high-confidence forecasts can raise MCC while pT falls [2]. No study reviewed here shows profit after fees, impact and latency [1, 2, 12].

For teams building market simulators. LOB-Bench gives distributional, discriminator and impact scores, plus a downstream test [7]. Generated data can reduce a forecaster's precision and recall, as one model's did in 2025 [7]. Inputs depend on how each book was rebuilt, from LOBSTER's message files [2, 6] to FI-2010's pre-processed release [1].

Rillor's market data is used for forecasting models tested out of sample on recorded history and scored against baselines, and for calibrating simulated markets from the spreads, depth and order flow of real books.

Limits of this analysis

- Horizons, labels, splits and normalizations differ across papers, so no score is compared across papers as one number.
- Most results come from U.S. and Nordic equities, plus London stocks [12] and bitcoin order books [19].
- LOBCAST's new datasets cover 10 trading days each for six stocks [1].
- LOBCAST is cited in its arXiv version of September 2023 [1].
- The 2019 test window in Briola, Bartolucci and Aste's Table 2 prints with the same dates as 2018 and starts on a Saturday, so only the 45, 5 and 10 day lengths are used [2].
- Kolm, Turiel and Westray's paper is cited from its abstract only [20].
- One LOB-Bench sentence gives a discriminator ROC score of 0.83 without naming the generator, so it is left out [7].
- No source reviewed reports net profit after fees, impact and latency, and this report gives no investment or trading advice.

Sources

1. LOB-Based Deep Learning Models for Stock Price Trend Prediction: A Benchmark Study. Prata, Masi, Berti and others. arXiv 2308.01915 v2, 19 September 2023, viewed 8 November 2026. https://arxiv.org/abs/2308.01915

2. Deep limit order book forecasting: a microstructural guide. Briola, Bartolucci and Aste. Quantitative Finance 25(7), 1101-1131, 22 July 2025, viewed 8 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC12315853/

3. Benchmark Dataset for Mid-Price Forecasting of Limit Order Book Data with Machine Learning Methods. Ntakaris, Magris, Kanniainen, Gabbouj and Iosifidis. arXiv 1705.03233 v5, 11 March 2020, viewed 8 November 2026. https://arxiv.org/abs/1705.03233

4. TLOB: A Novel Transformer Model with Dual Attention for Price Trend Prediction with Limit Order Book Data. Berti and Kasneci. arXiv 2502.15757 v3, 7 May 2025, viewed 8 November 2026. https://arxiv.org/abs/2502.15757

5. The Price Impact of Order Book Events. Cont, Kukanov and Stoikov. arXiv 1011.6402 v3, 13 April 2011, viewed 8 November 2026. https://arxiv.org/abs/1011.6402

6. The Short-Term Predictability of Returns in Order Book Markets: a Deep Learning Perspective. Lucchese, Pakkanen and Veraart. arXiv 2211.13777 v3, 8 October 2023, viewed 8 November 2026. https://arxiv.org/abs/2211.13777

7. LOB-Bench: Benchmarking Generative AI for Finance, an Application to Limit Order Book Data. Nagy, Frey, Li and others. arXiv 2502.09172 v2, 16 June 2025, viewed 8 November 2026. https://arxiv.org/abs/2502.09172

8. LOB-Bench (proceedings page). Nagy and others. Proceedings of the 42nd International Conference on Machine Learning, PMLR 267, 6 October 2025, viewed 8 November 2026. https://proceedings.mlr.press/v267/nagy25a.html

9. Nasdaq TotalView-ITCH 5.0 specification. Nasdaq. The Nasdaq Stock Market, revised to 28 April 2023, viewed 8 November 2026. https://www.nasdaqtrader.com/content/technicalsupport/specifications/dataproducts/NQTVITCHSpecification.pdf

10. Universal features of price formation in financial markets: perspectives from Deep Learning. Sirignano and Cont. arXiv 1803.06917, 19 March 2018, viewed 8 November 2026. https://arxiv.org/abs/1803.06917

11. Deep Learning for Limit Order Books. Sirignano. arXiv 1601.01987 v7, 5 July 2016, viewed 8 November 2026. https://arxiv.org/abs/1601.01987

12. DeepLOB: Deep Convolutional Neural Networks for Limit Order Books. Zhang, Zohren and Roberts. arXiv 1808.03668 v6, 23 January 2020 (IEEE Transactions on Signal Processing 67(11), 1 June 2019), viewed 8 November 2026. https://arxiv.org/abs/1808.03668

13. Transformers for Limit Order Books. Wallbridge. arXiv 2003.00130, 29 February 2020, viewed 8 November 2026. https://arxiv.org/abs/2003.00130

14. Deep Learning modeling of Limit Order Book: a comparative perspective. Briola, Turiel and Aste. arXiv 2007.07319 v3, 18 October 2020, viewed 8 November 2026. https://arxiv.org/abs/2007.07319

15. Benchmark Dataset for Mid-Price Forecasting of Limit Order Book Data with Machine Learning Methods (dataset record). Ntakaris and others. Fairdata metadata service, issued 1 June 2017, viewed 8 November 2026. https://metax.fairdata.fi/v3/datasets/73eb48d7-4dbc-4a10-a52a-da745b47a649

16. Deep Learning for Market by Order Data. Zhang, Lim and Zohren. arXiv 2102.08811 v2, 27 July 2021, viewed 8 November 2026. https://arxiv.org/abs/2102.08811

17. TensorLOB: Structure-Preserving Tensor Learning for Limit Order Book Forecasting. Gu and Mandic. arXiv 2611.02866 v1, 2 November 2026, viewed 8 November 2026. https://arxiv.org/abs/2611.02866

18. Deep Limit Order Book Forecasting. Briola, Bartolucci and Aste. arXiv 2403.09267 v4, 4 June 2024, viewed 8 November 2026. https://arxiv.org/abs/2403.09267

19. Exploring Microstructural Dynamics in Cryptocurrency Limit Order Books: Better Inputs Matter More Than Stacking Another Hidden Layer. Wang. arXiv 2506.05764 v2, 9 June 2025, viewed 8 November 2026. https://arxiv.org/abs/2506.05764

20. Deep order flow imbalance: Extracting alpha at multiple horizons from the limit order book. Kolm, Turiel and Westray. Mathematical Finance 33(4), 21 July 2023, viewed 8 November 2026. https://doi.org/10.1111/mafi.12413

21. Generative AI for End-to-End Limit Order Book Modelling: A Token-Level Autoregressive Generative Model of Message Flow Using a Deep State Space Network. Nagy, Frey, Sapora and others. arXiv 2309.00638, 23 August 2023, viewed 8 November 2026. https://arxiv.org/abs/2309.00638

22. Simulating and analyzing order book data: The queue-reactive model. Huang, Lehalle and Rosenbaum. arXiv 1312.0563 v2, 3 September 2014, viewed 8 November 2026. https://arxiv.org/abs/1312.0563

23. ABIDES: Towards High-Fidelity Market Simulation for AI Research. Byrd, Hybinette and Balch. arXiv 1904.12066, 26 April 2019, viewed 8 November 2026. https://arxiv.org/abs/1904.12066

24. Towards Realistic Market Simulations: a Generative Adversarial Networks Approach. Coletta, Prata, Conti and others. arXiv 2110.13287, 25 October 2021, viewed 8 November 2026. https://arxiv.org/abs/2110.13287

25. TRADES: Generating Realistic Market Simulations with Diffusion Models. Berti, Prenkaj and Velardi. arXiv 2502.07071 v3, 20 November 2025, viewed 8 November 2026. https://arxiv.org/abs/2502.07071

26. Painting the market: generative diffusion models for financial limit order book simulation and forecasting. Backhouse, Li, Foerster, Calinescu and Zohren. arXiv 2509.05107, 5 September 2025, viewed 8 November 2026. https://arxiv.org/abs/2509.05107

27. LOB-Impact Benchmark: Emergent Market Impact in Generative Limit Order Book Models. Nigmatulin, Mohl, Frey and others. arXiv 2610.27100, 19 October 2026, viewed 8 November 2026. https://arxiv.org/abs/2610.27100

28. DeepLOB-Deep-Convolutional-Neural-Networks-for-Limit-Order-Books (code). Zhang. GitHub, last pushed 15 July 2021, viewed 8 November 2026. https://github.com/zcakhaa/DeepLOB-Deep-Convolutional-Neural-Networks-for-Limit-Order-Books

29. LOBCAST (code). Prata and others. GitHub, last pushed 9 May 2024, viewed 8 November 2026. https://github.com/matteoprata/LOBCAST

30. LOBFrame (code). FinancialComputingUCL. GitHub, last pushed 31 May 2024, viewed 8 November 2026. https://github.com/FinancialComputingUCL/LOBFrame

31. lob_bench (code). peernagy. GitHub, last pushed 27 January 2026, viewed 8 November 2026. https://github.com/peernagy/lob_bench

About this report

Published by Rillor on 10 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Deep learning on limit order books: what benchmarks show. 10 November 2026. https://rillor.com/insights/deep-learning-limit-order-book-benchmarks

Notices

Rillor is not a registered investment adviser or commodity trading advisor and does not provide investment or trading advice. Research, forecasts, data and software described on this site are for research and engineering use. Nothing here is an offer or recommendation to buy or sell any security, commodity interest or digital asset. Past or simulated results do not indicate future results.

Work with Rillor

Send the instruments, the history window, the delivery format and the intended use. Contact us at rillor.com/contact?type=market-data or contact@rillor.com. More on market data: rillor.com/markets/market-data