УДК 336.7:004.8

Малышева Оксана

Финансовой аналитик

Opatra London LLP / Cosmo Tech LLP

Алматы, Казахстан

ORCID: https://orcid.org/0009-0005-0348-130X

ОЦЕНКА БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ФИНАНСОВОГО ПРОГНОЗИРОВАНИЯ: ТОЧНОСТЬ, НАДЕЖНОСТЬ И РИСК МОДЕЛИ

Аннотация. *Цель исследования — разработать проверяемую схему оценки больших языковых моделей (БЯМ), применяемых в финансовом прогнозировании. Проведён целевой сравнительный анализ восьми опубликованных работ о финансовых БЯМ, прогнозировании и валидации моделей; авторские эксперименты с временными рядами не проводились. Анализ даёт в качестве итога схему оценки по трём параметрам (риск модели, устойчивость к переменам условий и точность прогноза) и матрицу источников. Исследования показали: ставить в один ряд результаты, полученные при иных наборах данных, временных горизонтах или целевых переменных, не приходится; равно как и говорить о качестве БЯМ, опираясь на то, что дают гибридные модели без неё. В предлагаемой схеме заложены те минимальные данные, которые позволяют провести сравнение на воспроизводимой основе и увидеть, до каких пор можно доверять выводам из имеющейся литературы.*

Ключевые слова: *большие языковые модели; финансовое прогнозирование; точность прогноза; устойчивость; модельный риск; валидация.*

Oksana Malysheva
Financial Analytics
Opatra London LLP / Cosmo Tech LLP
Almaty, Kazakhstan
ORCID: https://orcid.org/0009-0005-0348-130X

EVALUATING LARGE LANGUAGE MODELS FOR FINANCIAL FORECASTING: ACCURACY, ROBUSTNESS, AND MODEL RISK

Abstract. *This study develops an auditable framework for evaluating large language models (LLMs) used in financial forecasting. We compared eight selected publications on financial LLMs, forecasting methods, and model validation; no original time-series experiment was conducted. The analysis produces a source matrix and a three-part framework addressing predictive accuracy, robustness under changing conditions, and model risk. Findings from studies with different targets, horizons, and datasets cannot be pooled into a common ranking, and results for non-LLM hybrid models do not measure LLM performance. The framework puts comparisons on a reproducible footing; it specifies the minimum required information and clarifies the boundaries of the evidence under review.*

Keywords: *large language models; financial forecasting; forecast accuracy; robustness; model risk; validation.*

1. INTRODUCTION

Financial institutions often use unstructured text, such as news or company disclosures, alongside numerical time series. A large language model may produce a classification or signal from this material, but careful reading of the text is no substitute for a foolproof forecast of the markets. For that, the measure of success, the benchmark, and what information is available when a decision is made must be put forward on their own terms.

Recent work shows both the difficulties and the potential. Lopez-Lira and Tang (2026) examined the predictability of stock returns from the news; Xie et al. (2024) then put financial LLMs through their paces on a variety of tasks and concluded that forecasting is no easy matter.

What sets an LLM apart from other predictive architectures is key. Even without a language model, the decomposition-and-LSTM system can deliver formidable forecasting. One might use the hybrid methods proposed by Dong and Zhou (2024) or Ozupek et al. (2024) as a baseline, but they should not be presented as LLM test results. FinGPT is an open-source framework that prepares financial text and data (Yang et al., 2023).

What then is needed to put an LLM forecast in finance to the test on its accuracy, model risk, and robustness? The challenge is how to make such an assessment without mixing evidence from one class of models or task with another. This paper contributes a source-to-claim matrix and a reporting framework that separates these three dimensions. It is a focused methodological analysis, not a claim that the reviewed studies prove one model class superior.

2. MATERIALS AND METHODS

2.1. Research design

We conducted a focused comparative review of a purposive set of eight publications. This review is not a systematic review or meta-analysis: we did not conduct exhaustive database search, pool effect estimates, or train independent models. Its unit of analysis is a published study and the particular claim that study can support. We distinguished three source roles before synthesis: direct LLM evaluation in finance; non-LLM forecasting comparators; and general methodological work on AI or financial model validation.

2.2. Source selection and extraction

The set contains three direct financial-LLM sources (Lopez-Lira & Tang, 2026; Xie et al., 2024; Yang et al., 2023), two non-LLM forecasting

comparators (Dong & Zhou, 2024; Ozupek et al., 2024), and three contextual or validation sources (Cao, 2022; Cummins et al., 2023; Remlinger et al., 2023). We selected sources that address at least one of the following: a financial LLM application, a financial forecasting benchmark, or the validation of predictive financial models. We excluded studies from unrelated application domains as evidence for financial forecasting. We extracted only claims relevant to the source's actual task.

For each publication, we extracted the following fields: model class, financial task, type of input, target forecast and horizon (if applicable), evaluation design, comparison baseline, robustness evidence, and model-risk implications. We used "Not established by this source" whenever the publication did not directly test a proposed claim. This prevents an application paper or a general discussion of AI from being misrepresented as a controlled LLM forecasting benchmark.

2.3. Evaluation procedure

Accuracy was defined relative to the specified prediction target. Directional stock-movement classification may use balanced accuracy or related classification measures; continuous price or return forecasts may use MAE or RMSE after disclosing the output scale. These metrics are examples of what a future experiment might report; we did not compute any new metric here. Benchmark comparison requires identical assets, observation periods, target definitions, decision timestamps, and forecast horizons.

We assessed robustness by how far a study reports performance under changes in sample period, market regime, text source, prompt formulation, or missing/noisy inputs. We assessed model risk through disclosures relevant to information leakage, model/data drift, interpretability, reproducibility, validation, and human oversight. We marked a source as offering direct evidence only for a condition it examined. The authors assigned no numeric robustness or risk scores.

2.4. Comparability and scope

The authors did not numerically compare reported errors across studies. A lower RMSE does not imply superiority if the series, units, horizon, or test windows differ. Financial sentiment classification, stock-return prediction, and a general financial question-answering benchmark are different tasks. Likewise, we used non-LLM hybrids only to specify candidate baselines. The literature selection is intentionally small; conclusions are restricted to the studies and task distinctions set out below.

3. RESULTS

3.1. Source-to-claim matrix

Table 1 records which role each source can legitimately play in the analysis. It reports no invented performance values. The classifications result from applying the source-role and task criteria described in Section 2.

Table 1. Role of the reviewed publications in evaluating LLM-based financial forecasting

Publication	Model and task	Permitted inference
Lopez-Lira & Tang (2026).	Stock return and reaction prediction from the news	The evidence for LLM forecasting is limited to the news task at hand; it does not serve as a market benchmark of any kind.
Xie et al. (2024).	Forecasting and other financial LLM work with FinBen	You will find a direct cross-task LLM benchmark here, but do not conflate stock-return studies with task-specific scores.
Yang et al. (2023)	The FinGPT framework for an LLM in finance, together with its data pipeline	As for architecture and data curation, there is no general case for better forecast returns.
Dong & Zhou (2024); Ozupek et al. (2024)	Time-series models of a non-LLM hybrid nature	Some comparators may be used in forecasting, yet their results are not due to the LLMs.

Publication	Model and task	Permitted inference
Cao (2022); Cummins et al. (2023)	Model validation in the financial sector via AI	What is offered is context on risk controls and deployment, not a straight test of LLM forecasts.
Remlinger et al. (2023).	The use of expert aggregation to forecast in finance	In short, a concept for comparison, nothing more than an evaluation of the LLM.

One can see a category error in the manuscript as it was previously put forward. In that iteration, hybrid time-series models were put on a par with LLMs, and numerical scores unrelated to one another were juxtaposed. Do away with them and what is left is a more circumscribed result, yet one that can be audited; it offers direct evidence for some LLM financial tasks where there is no basis for a common ranking of every architecture under review.

3.2. Accuracy and task definition

In their 2026 paper, Lopez-Lira and Tang examine the stock market's reaction to firm-specific news headlines as an LLM would. They define the task as a text-based prediction exercise. Then Xie et al. (2024) put 21 LLMs through their paces on 42 datasets and 24 financial tasks; they find LLMs do not fare as well on forecasting and other complex work as they do on information extraction or textual analysis. One need not see these as at odds, since the scoring, data, and targets differ. Neither study authorizes importing an error value from a hybrid price-series model as a score for ChatGPT or FinGPT.

A valid accuracy claim therefore needs four labels next to any numerical result: the target (for example, direction rather than price), horizon, test period, and baseline. It should also state whether an LLM produced the output directly or whether a separate forecasting model used LLM-derived text features. The reviewed sources establish the necessity of this distinction; the present article does not estimate a new accuracy rate.

3.3. Robustness and risk controls

The rapid evolution of financial information demands a current, well-constructed data pipeline, as Yang et al. (2023) note in their FinGPT study. As Xie et al. (2024) argue, assessing performance on any one task is insufficient; after all, the capabilities tested by language understanding are not the same as those for decision-making or forecasting. Sensitivity checks on inputs and prompts, out-of-period evaluation, and open handling of new market conditions are needed. This article puts forward such measures as requirements rather than stress tests.

One must also be mindful of uncertainty in model parameters and their distribution when validating a financial model (Cummins et al., 2023). When the forecast is LLM-assisted, an additional practical control is auditing the information at hand. It is incumbent that every piece of data, be it an article, filing, or a feature derived from the market, was in existence prior to the time of the simulated decision. Any impressive outcome hinging on information made public after the fact would not prove a model's utility for forecasting. Human review and documented model updates remain relevant when an output supports a consequential financial decision (Cao, 2022).

3.4. Derived evaluation framework

The synthesis yields a three-stage reporting framework. The study first defines the financial decision and measures accuracy against a same-task baseline. It next probes robustness across periods, regimes, prompts, and data quality. Finally, it documents leakage checks, validation, drift monitoring, reproducibility, and oversight. A model should not receive a composite score where one dimension was not measured; the missing dimension should be disclosed. Table 2 translates this framework into a reproducible reporting checklist.

Table 2. Minimum information for a controlled future LLM forecasting comparison

Dimension	Required report	Reason
Accuracy	Target; units; horizon; chronological train/test split; baseline; metric; uncertainty	Ensures reported gains apply to the same decision and test data.
Robustness	Market periods; volatility regimes; prompt variants; text-source changes; missing/noisy input tests	Reveals whether performance holds up under reasonable changes in use conditions.
Model risk	Timestamp audit; model/version record; leakage checks; drift monitoring; human review	Makes forecast provenance and operational limits inspectable.

4. DISCUSSION

The source matrix changes how we interpret literature. Evidence that a particular LLM extracts useful information from headlines does not establish that every LLM reliably forecasts prices. Evidence that a non-LLM hybrid predicts a financial series well does not establish the value of language modeling. Both are relevant to designing a future head-to-head test, but they occupy different positions in the causal chain: text interpretation may create features, while a forecasting system maps available features to a target.

The proposed framework also explains why accuracy alone is incomplete. Even when a reported metric improves, a deployable method must remain useful when the information environment changes and must be evaluated without future information leaking into past predictions. The robustness and risk columns are therefore separate reporting obligations; not arbitrary penalties attached to an accuracy score. No threshold or weighting is proposed because the acceptable trade-off depends on the financial decision, the cost of an error, and whether a human can intervene.

The study has three limitations. Selection was purposive rather than exhaustive; some relevant financial LLM evaluations may not appear in the matrix. The authors did not re-estimate full numerical performance on a shared dataset. The framework is conceptual and has not been validated by a prospective user study or controlled backtest. It should be treated as an evaluation protocol for future work, not proof that a particular model is safe or profitable.

5. CONCLUSION

A credible evaluation of LLMs for financial forecasting must define the target, data availability, horizon, baseline, and validation period before it reports predictive accuracy. The reviewed literature contains direct LLM evidence for specific financial tasks and useful non-LLM comparators, but it does not support a pooled ranking of the models discussed in this article. The original contribution is a source-to-claim matrix and a three-stage reporting framework that couples accuracy with robustness and model-risk checks.

Future research should test the named LLM and non-LLM baselines on the same time-stamped financial dataset, using a chronological holdout and pre-specified measures. It should repeat the evaluation across market regimes and disclose prompt sensitivity, information leakage checks, and model updates. Until those conditions are met, limit comparative claims to the exact task and evidence each source reports.

REFERENCES

1. Cao, L. (2022). AI in finance: Challenges, techniques, and opportunities. ACM Computing Surveys, 55(3), Article 64. DOI: 10.1145/3502289.
2. Cummins, M., Gogolin, F., Kearney, F., Kiely, G., & Murphy, B. (2023). Practice-relevant model validation: Distributional parameter risk analysis

in financial model risk management. Annals of Operations Research, 330, 431–455. DOI: 10.1007/s10479-022-04574-x.

3. Dong, Z., & Zhou, Y. (2024). A novel hybrid model for financial forecasting based on CEEMDAN-SE and ARIMA-CNN-LSTM. Mathematics, 12(16), Article 2434. DOI: 10.3390/math12162434.
4. Lopez-Lira, A., & Tang, Y. (2026). Can ChatGPT forecast stock price movements? Return predictability and large language models. Journal of Financial Economics, Article 104335. DOI: 10.1016/j.jfineco.2026.104335.
5. Ozupek, O., Yilmaz, R., Ghasemkhani, B., Birant, D., & Kut, R. A. (2024). A novel hybrid model (EMD-TI-LSTM) for enhanced financial forecasting with machine learning. Mathematics, 12(17), Article 2794. DOI: 10.3390/math12172794.
6. Remlinger, C., Alasseur, C., Brière, M., & Mikael, J. (2023). Expert aggregation for financial forecasting. Journal of Finance and Data Science, 9, Article 100108. DOI: 10.1016/j.jfds.2023.100108.
7. Xie, Q., Han, W., Chen, Z., et al. (2024). FinBen: A holistic financial benchmark for large language models. Advances in Neural Information Processing Systems, 37, Datasets and Benchmarks Track. DOI: 10.52202/079017-3033.
8. Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-source financial large language models. arXiv:2306.06031. DOI: 10.48550/arXiv.2306.06031.