

AI-Augmented Strategic Forecasting: Integrating Large Language Models with Traditional Business Intelligence Systems

Shuvam Kumar Das

Independent Researcher, Mountain House, California, United States

Abstract—Strategic forecasting in enterprises is commonly implemented as a business-intelligence (BI) workflow over structured transactional data, while strategic signals contained in reports, event descriptions, customer communications, and other unstructured sources remain difficult to incorporate systematically. This paper proposes an AI-Augmented Strategic Forecasting (AISF) architecture that preserves quantitative forecasting models as the numerical prediction layer and introduces retrieval augmented generation (RAG) and large language models (LLMs) as a contextual intelligence layer. The empirical design uses the publicly available M5 Walmart retail-demand benchmark, which contains hierarchical daily unit-sales series, prices, calendar information, and event variables. The study compares ARIMA, Prophet, XGBoost, and LSTM baselines with an LLM augmented hybrid in which retrieved event context is converted into structured qualitative signals and fused with the quantitative forecast. Evaluation uses rolling-origin back testing, MAE, RMSE, and MASE, together with evidence-grounding and scenario quality measures. The paper also defines ablations that remove retrieval or qualitative signal extraction. The resulting framework is designed to test whether LLM augmentation improves not only numerical accuracy but also the interpretability and strategic usefulness of demand forecasts, without allowing generated text to override auditable quantitative evidence.

Keywords—artificial intelligence, business intelligence, demand forecasting, large language models, retrieval-augmented generation, time series forecasting.

I. INTRODUCTION

Enterprise forecasting systems have historically separated quantitative prediction from managerial interpretation. Enterprise resource planning (ERP), customer relationship management (CRM), and data-warehouse systems provide structured observations that feed dashboards, statistical models, and machine-learning pipelines. These systems are effective for extracting seasonality, trends, promotions, prices, and other measurable drivers, but they are less effective at incorporating qualitative context such as event descriptions, management commentary, product announcements, or textual explanations of unusual demand.

Recent work on large language models has opened a complementary route. LLMs can summarize heterogeneous information, identify relationships expressed in natural language, and support open-ended decision processes. Handler *et al.* argue that this versatility changes the design space for decision support systems, but also introduces new questions concerning evidence, organizational use, and human judgment [1]. In time-series research, recent studies show that pretrained language-model architectures can be adapted to forecasting, including zero-shot numerical forecasting and explicit time-series reprogramming [14], [16]. A 2026 systematic review further identifies rapid growth in LLM-based time-series forecasting while emphasizing unresolved issues in representation alignment and evaluation [2].

This paper focuses on a narrower enterprise problem: how to augment an existing BI forecasting pipeline without replacing the auditable numerical forecasting layer. The proposed AISF architecture therefore assigns distinct roles to quantitative models and LLMs. Quantitative models generate the primary forecast; RAG retrieves time-aligned evidence; the LLM extracts qualitative signals and produces explanations; and a scenario engine combines these outputs under explicit assumptions. This separation is intended to reduce unsupported claims and make the system compatible with existing governance controls.

This study makes four contributions. First, it formalizes an enterprise architecture connecting operational data, BI, forecasting models, RAG, LLM analysis, scenario generation, and executive decision support. Second, it defines a reproducible empirical protocol on the M5 retail sales benchmark. Third, it compares classical statistical, machine learning, deep learning, and LLM augmented configurations under chronological evaluation. Fourth, it introduces grounding and scenario quality measures to forecast error, recognizing that strategic forecasting quality is multidimensional.

II. RELATED WORK

A. Business Intelligence and Forecasting

Forecasting research has established that no single model dominates across all demand structures. The M4 competition demonstrated the importance of empirical benchmarking across heterogeneous time series and showed the value of combinations and strong statistical baselines [5]. The M5 competition extended this setting to hierarchical retail demand with prices, calendar variables, and intermittency [3]. Subsequent analyses found strong performance from gradient boosted trees and neural networks on the M5 task [4]. Forecast evaluation also requires care: Hyndman and Koehler recommend scaled measures such as MASE when comparing series with different scales [6].

Classical and modern forecasting models remain important components of enterprise systems. Prophet provides an additive framework designed for business timeseries with multiple seasonalities and events [7]. DeepAR demonstrates the value of recurrent probabilistic forecasting across related series [10]. LSTM networks remain a widely used sequence model because their gating mechanism can preserve long range information [9]. XGBoost provides a scalable tree boosting implementation suitable for lagged demand, price, calendar, and promotion features [8]. Transformer-based architectures such as Informer, Autoformer, and the Temporal Fusion Transformer extend this family toward long-range dependencies, decomposition, multi-horizon forecasting, and interpretable covariate selection [11]–[13].

B. Foundation Models for Time Series

A second literature stream investigates whether pretrained language-model mechanisms can generalize to numerical sequences. Gruver et al. show that LLMs can be prompted as zero-shot time-series forecasters by representing numerical values as tokens [14]. Zhou et al. investigate pretrained language models for general time-series analysis [15]. Time-LLM reprograms time-series patches into a representation that can be processed by a frozen LLM [16]. Chronos instead tokenizes scaled and quantized time-series values and trains transformer language-model architectures directly on time-series data [17]; TimesFM explores a decoder-only foundation model trained on a large corpus of time-series points [18]. PatchTST and related transformer approaches further demonstrate that suitable representations can substantially improve long-horizon forecasting [19].

These studies, however, do not fully address the enterprise integration problem. In an operational BI setting, the question is not only whether an LLM can predict a number, but whether it can use auditable business evidence to improve an established forecast and communicate implications to decision makers. This motivates the hybrid architecture in this paper.

C. Retrieval-Augmented Generation and Decision Support

Retrieval-augmented generation combines external retrieval with generation so that model outputs can be conditioned on a selected evidence set rather than relying only on parametric memory [20]. For enterprise forecasting, this is important because the contextual evidence may change after model training and must be traceable to a source and date. Handler et al. similarly emphasize that LLM enabled decision support creates new questions about evidence collection, alternatives, and choice [1]. AISF adopts this principle by making retrieved evidence an explicit input to the strategic interpretation layer.

III. AISF ARCHITECTURE

Fig. 1 shows the proposed architecture. The architecture deliberately separates deterministic data processing from probabilistic generation. ERP, CRM, data-warehouse, and external business event sources feed a governed data layer. BI transforms these records into analytical features and dashboards. Forecasting models produce a quantitative forecast and uncertainty information. A retrieval service indexes time-stamped contextual documents. The LLM converts retrieved documents into structured qualitative signals, each associated with evidence. The scenario engine then evaluates baseline, optimistic, pessimistic, and disruption cases before an executive decision layer presents recommendations, assumptions, and citations.

The core design principle is that the LLM does not directly overwrite the numerical forecast. Let $\hat{y}^{(q)}_{t+h}$ denote the quantitative forecast and let c_t denote a vector of extracted contextual signals. The strategic output is

$$O_{t+h} = g(\hat{y}^{(q)}_{t+h}, c_t, S, A, \quad (1)$$

where S is a set of scenarios and A is a set of explicit assumptions. The architecture therefore supports traceability from an executive recommendation back to both a numerical model and its textual evidence.

IV. METHODOLOGY AND EXPERIMENTAL DESIGN

A. Dataset

The primary benchmark is the M5 Walmart Forecasting Accuracy dataset. The benchmark contains daily unit sales for Walmart products across stores in California, Texas, and Wisconsin, together with calendar information and selling prices. The competition defines 42,840 hierarchical series and a 28-day forecasting task [3]. The dataset is publicly distributed through the M5 competition platform [22].

For a reproducible conference-scale experiment, the study uses a stratified subset rather than all 42,840 series. The proposed subset contains 30 representative item-store series sampled across states and product categories, with at least 1,500 historical observations. Each series is aligned to the calendar table. The textual-context corpus is constructed from the human-readable event and event-type fields in the calendar data, preserving the original event date and event category. This choice does not claim that M5 contains free-form news; rather, it tests whether an LLM can transform structured event descriptions into a contextual signal that is useful to an existing demand forecast. A secondary extension can replace this corpus with contemporaneous news or earnings-call text while preserving the same evaluation protocol.

B. Data Preparation and Leakage Control

The data are divided chronologically into training, validation, and test periods. No observation, event description, or price information occurring after the forecast origin is permitted in the input. Hyperparameters are selected on validation windows and frozen before final testing. Rolling origin evaluation is used because a single holdout can make results sensitive to one particular demand regime. For each forecast origin t , the quantitative model receives the feature matrix $X_{\leq t}$ and predicts

$$\hat{y}_{t+h} = f(y_{\leq t}, X_{\leq t}, h). \quad (2)$$

The RAG component receives a query containing the forecast origin, product/store identifiers, and planning horizon. It retrieves only documents whose timestamps satisfy the information availability constraint. The LLM then maps the retrieved evidence into structured signals

$$c_i = (\text{type}_i, \text{direction}_i, \text{strength}_i, \text{evidence}_i). \quad (3)$$

Signals are stored separately from generated prose so that downstream models can consume auditable features.

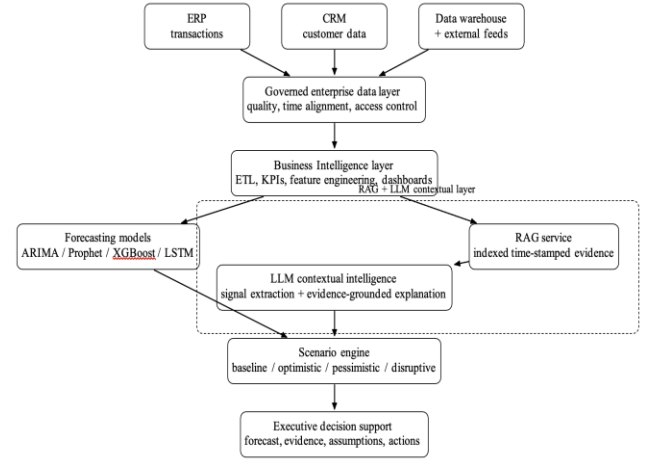


Fig. 1. AI-Augmented Strategic Forecasting (AISF) architecture. Quantitative forecasts remain the primary numerical evidence; RAG and the LLM add time-aligned contextual signals and explanations.

TABLE I
EXPERIMENTAL DATASET AND INPUTS

Component	Description
Target series	Daily unit sales by item-store
History	Up to 1,941 daily observations per series
Structured inputs	Lagged sales, rolling statistics, price,
Context	weekday, month, SNAP indicators Date-aligned event names, event types, and calendar descriptions
Forecast horizons	1, 7, and 28 days
Evaluation windows	Rolling-origin test

C. Baseline Models

Four forecasting baselines are included. (1) ARIMA represents a classical statistical baseline and is selected using training-window diagnostics. (2) Prophet models trend, multiple seasonal components, and event effects [7]. (3) XGBoost uses lag, rolling, calendar, price, and promotion feature and is trained on supervised sliding windows [8]. (4) LSTM uses a fixed look-back sequence and predicts each horizon with a multi-output head [9]. These baselines span statistical, additive, tree-based, and recurrent approaches.

D. LLM-Augmented Model

The proposed AISF model is a two-stage hybrid. First, XG-Boost produces the numerical demand forecast. Second, RAG retrieves time-aligned event context and an LLM converts that context into a compact qualitative representation. The final fusion model combines the numerical forecast with contextual features rather than asking the LLM to invent the numerical trajectory:

$$\hat{y}_{t+h}^{AISF} = F(\hat{y}_{t+h}^{XGB}, c_t, X_t). \quad (4)$$

The fusion function F is implemented as a lightweight gradient-boosted regressor trained only on the training period. This preserves a deterministic numerical path and makes it possible to compare the marginal contribution of contextual signals.

The LLM prompt is fixed across all test cases. It requires the model to: (a) identify events relevant to demand; (b) assign direction and confidence; (c) quote or cite the retrieved evidence identifier; (d) distinguish observed facts from assumptions; and (e) avoid changing the quantitative forecast unless the fusion model receives the extracted signal. Temperature is fixed, retrieval top- k is fixed, and repeated runs are used to measure output stability.

E. Ablation Study

Four configurations isolate the contribution of each component:

- 1) BI-only: quantitative forecasting without LLM context.
- 2) LLM-only: direct LLM forecast from serialized numerical history and context.
- 3) Hybrid-no-RAG: quantitative forecast plus LLM analysis without external retrieval.
- 4) Full AISF: quantitative forecast plus RAG-grounded LLM signals and fusion.

This design tests whether any observed benefit arises from the LLM itself, from retrieval, or from the structured hybrid architecture.

F. Evaluation Metrics

Forecast accuracy is evaluated using MAE, RMSE, and MASE. For a series y , MAE and RMSE are

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, \quad (5)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}. \quad (6)$$

MASE follows Hyndman and Koehler [6]. Relative improvement is calculated against the ARIMA and XGBoost baselines. Paired forecast errors across rolling origins are compared with the Diebold–Mariano test [21].

Because numerical accuracy alone does not establish strategic usefulness, two additional metrics are reported. Evidence-grounding score is

$$GS = \frac{1}{N} \sum_{i=1}^N g_i, \quad (7)$$

where $g_i = 1$ when a material factual claim is supported by retrieved evidence and 0 otherwise. Scenario quality is scored from 1–5 by blinded evaluators on plausibility, evidence support, distinctiveness, strategic relevance, assumption clarity, and actionability. Inter-rater reliability is reported using an appropriate agreement statistic.

V. EXPERIMENTAL PROTOCOL

The primary comparison is conducted over the same rolling-origin windows for all models. The principal horizon is 28 days, matching the M5 competition, with 1- and 7-day horizons included to test whether contextual information is more useful for near-term or longer planning. Each model receives the same structured variables where applicable. The LLM receives only information available at the forecast origin.

TABLE II
MODEL CONFIGURATIONS

ID	Model	Main inputs
M1	ARIMA	Sales history
M2	Prophet	Sales + calendar events
M3	XGBoost	Lags + price + calendar + events
M4	LSTM	Sales + structured covariates
M5	LLM-only	Serialized sales + context
M6	Hybrid-no-RAG	Quantitative forecast + generated context
M7	AISF	XGBoost + RAG + LLM signals

TABLE III
FORECASTING PERFORMANCE ON M5 TEST WINDOWS

Model	MAE	RMSE	MASE	Δ vs. XGB
ARIMA	TBD	TBD	TBD	TBD
Prophet	TBD	TBD	TBD	TBD
XGBoost	TBD	TBD	TBD	–
LSTM	TBD	TBD	TBD	TBD
LLM-only	TBD	TBD	TBD	TBD
Hybrid-no-RAG	TBD	TBD	TBD	TBD
AISF	TBD	TBD	TBD	TBD

Table II summarizes the planned comparison. The experiment is considered successful only if the full AISF configuration is evaluated on the identical test windows and if all hyperparameters, prompt templates, retrieval settings, and random seeds are recorded.

VI. RESULTS REPORTING PLAN

No numerical result is fabricated in this manuscript. Before submission, the following tables should be populated directly from the reproducible experiment. This distinction is essential because IEEE conference guidance requires results to report the work actually performed and interpreted rather than anticipated findings [26].

A second table should report grounding rate, unsupported claims, scenario quality, and decision time. The paper should report confidence intervals or paired statistical tests where appropriate. If AISF improves explanation quality but not numerical accuracy, that result should be retained rather than treated as a failure: it would indicate that LLM augmentation primarily adds contextual decision-support value. Conversely, if accuracy improves but grounding decreases, the architecture would require stronger retrieval and governance controls.

VII. GOVERNANCE AND THREATS TO VALIDITY

The architecture introduces risks not present in conventional forecasting. LLM outputs can contain unsupported statements, retrieval can surface temporally invalid evidence, and generated scenarios can be persuasive even when the underlying quantitative forecast is uncertain. The proposed controls therefore include source timestamps, evidence identifiers, fixed retrieval policies, human review of material recommendations, prompt/version logging, and monitoring for drift.

External validity is limited by the M5 benchmark. Although M5 is a realistic retail forecasting environment and has been widely studied, its calendar context is not equivalent to enterprise CRM text, earnings calls, or customer-service records.

A second limitation is model dependence: different LLMs and prompts may produce different qualitative signals. Repeated runs and frozen prompts reduce but do not eliminate this issue. Finally, the proposed LLM-only baseline is intentionally included as a stress test; it should not be interpreted as a replacement for specialized forecasting models.

VIII. CONCLUSION

This paper presents an empirical architecture for integrating LLMs into established business-intelligence forecasting systems without making generative models the sole source of numerical prediction. The AISF design assigns quantitative forecasting to auditable statistical and machine-learning models and uses RAG and LLMs to transform time-aligned contextual evidence into structured qualitative signals. The M5 Walmart benchmark provides a reproducible starting point for comparing ARIMA, Prophet, XGBoost, LSTM, LLM-only, and hybrid configurations under rolling-origin evaluation.

The central empirical question is therefore not whether an LLM can forecast sales in isolation, but whether evidence-grounded language intelligence adds measurable value to an existing BI forecasting pipeline. The proposed combination of forecast accuracy, evidence grounding, scenario quality, and ablation analysis is intended to answer that question while preserving reproducibility and governance. The next stage of the work is to execute the protocol, populate the results tables, and validate whether the observed gains justify the additional complexity of the AISF architecture.

Acknowledgment

The authors acknowledge the maintainers of the M5 Forecasting Accuracy benchmark and the open research community whose public forecasting datasets and implementations support reproducible evaluation.

REFERENCES

- [1] A. K. Handler, K. R. Larsen, and R. D. Hackathorn, “Large language models present new questions for decision support,” *International Journal of Information Management*, vol. 79, Art. no. 102811, 2024, doi: 10.1016/j.ijinfomgt.2024.102811.
- [2] G. I. F. de Paiva, A. C. V. e Pinto, M. A. Alves, O. Orang, L. M. Astore, M. V. M. Oliveira, and F. G. Guimarães, “Leveraging large language models for time series forecasting: A systematic literature review,” *Knowledge-Based Systems*, vol. 343, Art. no. 115938, 2026, doi: 10.1016/j.knosys.2026.115938.
- [3] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “The M5 competition: Background, organization, and implementation,” *International Journal of Forecasting*, vol. 38, no. 4, pp. 1325–1336, 2022, doi: 10.1016/j.ijforecast.2021.07.007.

- [4] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1346–1364, 2022, doi: 10.1016/j.ijforecast.2021.11.013.
- [5] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, vol. 36, no. 1, pp. 54–74, 2020, doi: 10.1016/j.ijforecast.2019.04.014.
- [6] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006, doi: 10.1016/j.ijforecast.2006.03.001.
- [7] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018, doi: 10.1080/00031305.2017.1380080.
- [8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [10] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020, doi: 10.1016/j.ijforecast.2019.07.001.
- [11] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient Transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 11106–11115, doi: 10.1609/aaai.v35i12.17325.
- [12] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition Transformers with auto-correlation for long-term series forecasting," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [13] B. Lim, S. O. Ank, N. Loeff, and T. Pfister, "Temporal fusion Transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.
- [14] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson, "Large language models are zero-shot time series forecasters," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, doi: 10.52202/075280-0861.
- [15] T. Zhou, P. Niu, X. Wang, L. Sun, and R. Jin, "One Fits All: Power general time series analysis by pretrained LM," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 43322–43355, doi: 10.52202/075280-1877.
- [16] M. Jin, S. Wang, L. Ma, Z. Chu, J. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, and Q. Wen, "Time-LLM: Time series forecasting by reprogramming large language models," in *Proc. Int. Conf. Learning Representations*, 2024.
- [17] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shehur, S. S. Rangapuram, S. Pineda Arango, S. Kapoor, J. Zschiegner, C. Maddix, H. Wang, M. W. Mahoney, K. Torkkola, A. G. Wilson, M. Bohlke-Schneider, and Y. Wang, "Chronos: Learning the language of time series," *Transactions on Machine Learning Research*, 2024.
- [18] R. Sen, Y. Zhou, and colleagues, "A decoder-only foundation model for time-series forecasting," in *Proc. 41st Int. Conf. Machine Learning*, 2024.
- [19] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with Transformers," in *Proc. Int. Conf. Learning Representations*, 2023.
- [20] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.
- [21] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 253–263, 1995, doi: 10.1080/07350015.1995.10524599.
- [22] University of Nicosia, "M5 Forecasting–Accuracy," Kaggle, 2020. [Online]. Available: <https://www.kaggle.com/competitions/m5-forecasting-accuracy/data>
- [23] C. Chatfield, "The Holt-Winters forecasting procedure," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 27, no. 3, pp. 264–279, 1978, doi: 10.2307/2347162.
- [24] B. Seaman and J. Bowman, "Applicability of the M5 to forecasting at Walmart," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1468–1472, 2022, doi: 10.1016/j.ijforecast.2021.06.002.
- [25] E. Theodorou, S. Wang, Y. Kang, E. Spiliotis, S. Makridakis, and V. Assimakopoulos, "Exploring the representativeness of the M5 competition data," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1500–1506, 2022, doi: 10.1016/j.ijforecast.2021.07.006.