Artificial Intelligence and Language Models

ACS Publishing

Article

Leveraging Large Language Models for Indonesian Retail Sales Probabilistic Forecasting

M Dzaky Efendi [1], Salhazan Nasution [2], Mondheera Pituxcoosuvarn [3]

[1] Department of Informatics Engineering, Universitas Islam Riau, Indonesia
[2] Department of Informatics, Faculty of Engineering, University of Riau, Indonesia
[3] Faculty of Information Science and Engineering, Ritsumeikan University, Japan
* Correspondence: efendidzakyy@gmail.com;

Abstract: This research evaluates the effectiveness of Large Language Models (LLMs) for probabilistic forecasting of the Indonesian Retail Sales Index. We analyze monthly retail sales index data from Bank Indonesia, spanning January 2012 to January 2025 across seven product categories. A broad spectrum of time series forecasting models is developed using AutoGluon Time Series, including a baseline seasonal naive model, machine learning-based tabular models, classical statistical models (AutoETS, Dynamic Optimized Theta, and NPTS), deep learning models (Temporal Fusion Transformers, PatchTST, TiDE, and DeepAR), and transformer-based LLMs from the Chronos and Chronos Bolt families. For the LLM models, we consider both zero-shot forecasting (direct application of pre-trained models) and fine-tuning on the historical retail data. All models are evaluated on a hold-out test period using seven metrics: Scaled Quantile Loss (SQL), Weighted Quantile Loss (WQL), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Symmetric MAPE (SMAPE). The fine-tuned Chronos [base] model achieved the best overall performance, yielding the lowest errors with SQL = 0.274, WQL = 0.136, MAE = 0.184, MAPE = 0.267, MSE = 0.059, RMSE = 0.243, and SMAPE = 0.218. These results highlight the potential of LLM-based models to improve the accuracy of retail sales forecasts in Indonesia, especially in capturing long-term trends, while underscoring the remaining challenges in modeling short-term fluctuations.

Keywords: Time Series Forecasting, Chronos, Large Language Models, AutoGluon, Retail Sales Index, Probabilistic Forecasting, Indonesia

Citation: M Dzaky Efendi, Salhazan Nasution, Mondheera Pituxcoosuvarn. Leveraging Large Language Models for Indonesian Retail Sales Probabilistic Forecasting. *Artificial Intelligence and Language Models* 1–18. https://doi.org/

Received: 16 March 2026
Revised: 30 April 2026
Accepted: 14 May 2026
Published: 16 May 2026

1. Introduction

Precise retail sales forecasting plays a vital role in shaping successful business strategies and robust economic planning [1]. From a broader economic viewpoint, retail sales data serve as a key measure of consumer expenditure, offering insights into the economy's overall condition and valuable information for policymakers [2]. The ability to predict future retail trends accurately is therefore of paramount importance for decision-making at both the micro level (e.g., individual businesses) and the macro level (e.g., government policy).

Historically, time series forecasting has predominantly relied on statistical techniques such as ARIMA, exponential smoothing, and regression-based approaches [3–5]. While these methods have proven valuable in many contexts, they often struggle to capture the complex non-linear patterns and long-range dependencies that characterize modern economic data, including retail sales. The increasing availability of large datasets and advancements in computational power have paved the way for more sophisticated techniques, including machine learning and deep learning. Approaches such as Support Vector Regression (SVR)[6], Random Forests[7], and various neural network architectures (e.g., Re-

current Neural Networks and Convolutional Neural Networks) have shown encouraging results in improving forecasting accuracy across different domains [8,9].

In recent years, a paradigm shift has occurred in the field of artificial intelligence with the emergence of Large Language Models (LLMs), which are transformer-based architectures originally designed for natural language processing [10–12]. These architectures have exhibited impressive abilities in comprehending, generating, and reasoning with sequential data across multiple domains [13]. Their capacity to learn intricate patterns and contextual relationships from vast amounts of text data has sparked interest in exploring their potential for time series analysis and forecasting [14]. Although the use of LLMs in time series analysis is still an emerging area, early research suggests their promise in modeling complex temporal patterns and potentially surpassing conventional methods in specific contexts [15].

The retail sector in Indonesia constitutes a major part of the national economy, characterized by dynamic consumer behavior and vulnerability to various internal and external influences (such as seasonal festivities, economic fluctuations, and cultural events) [16,17]. Accurate forecasting of retail sales in this context can provide invaluable insights for businesses (for inventory management, marketing, and supply chain optimization) and for government agencies in formulating economic policies[18,19]. Despite the importance of this task, there is limited research specifically focused on leveraging cutting-edge LLMs for probabilistic forecasting of Indonesian retail sales.

This research addresses the above gap by investigating the effectiveness of several prominent LLMs, facilitated by the AutoGluon Time Series framework, for forecasting the Indonesian Retail Sales Index across multiple product categories. We compare the performance of these LLMs, in both zero-shot and fine-tuned settings, against a diverse set of traditional forecasting models, including baseline, statistical, tabular machine learning, and deep learning approaches. By conducting a comprehensive evaluation with multiple accuracy metrics, this study provides insights into the potential of LLMs to enhance forecast accuracy and probabilistic estimation for retail sales in the Indonesian context, ultimately contributing to the advancement of forecasting methodologies in this critical economic sector.

2. Materials and Methods

2.1. Data

This study employs historical monthly Retail Sales Index data sourced from Bank Indonesia, covering the period from January 2012 to January 2025 [20]. The dataset encompasses seven distinct product categories: *Spare Parts and Accessories*, *Food, Beverages & Tobacco*, *Motor Vehicle Fuel*, *Information and Communication Equipment*, *Other Household Appliances*, *Cultural and Recreational Goods*, and *Other Goods*. These categories represent a significant portion of the Indonesian retail market, providing a comprehensive view of consumer spending trends across different sectors. For the purpose of model training and evaluation, we treat each product category as a separate time series. We reserve the last 12 months of each series (February 2024 to January 2025) as the out-of-sample test set for evaluating forecast performance, while data up to January 2024 are used for training the models.

2.2. Data Pre-processing

The initial data was obtained from a comprehensive Excel file downloaded from the official Bank Indonesia website [20]. The original file comprised several sheets presenting the retail sales index at different temporal resolutions, including daily, monthly, and yearly frequencies. For the purposes of this study, only the monthly data was deemed relevant. Consequently, the initial step involved manually identifying and isolating the sheet corresponding to the monthly retail sales index. This data was subsequently extracted and stored in a separate Comma-Separated Values (CSV) file to facilitate a more efficient and focused modeling process.

A comprehensive evaluation of data quality was subsequently performed on the curated dataset. The analysis verified that all entries corresponding to the seven product categories over the defined time span were intact and complete. There were no missing values identified, nor were any irregular or outlier data points observed that would necessitate removal or the use of corrective methods such as imputation. As a result, the dataset was confirmed to be clean and fully prepared for direct application in time-series modeling.

To standardize the input data, we applied z-score normalization, a standard technique in machine learning preprocessing [21,22]. This method transforms the data for each time series to have a mean of zero and a standard deviation of one. This process is crucial for enhancing the performance and stability of many forecasting models, particularly those sensitive to the scale of input features, by ensuring all series are on a comparable scale [23]. Table 1 provides a sample of the data values for two categories before and after this transformation. After the forecasting process was complete, all predicted values were de-normalized back to the original scale of the Retail Sales Index to allow for meaningful evaluation and interpretation of the results.

Table 1. Comparison of sample Retail Sales Index (RSI) values before and after z-score normalization.

Month	Before Normalization	After Normalization
January 2012	95.5	-1.4083
February 2012	86.9	-1.9284
March 2012	93.0	-1.5595
April 2012	86.9	-1.9284
May 2012	93.1	-1.5535

2.3. Methods

In this study, we evaluate a diverse set of time series forecasting models available in the **AutoGluon-TimeSeries** Python library. To ensure a fair and equivalent comparison, the hyperparameters for all models were manually configured to align the training setups. This allows performance differences to be attributed more directly to the model architectures themselves. The models are grouped into five categories. The specific parameters for all non-LLM models are consolidated in Table 2 for clarity.

2.3.1. Baseline Model

A simple baseline model is used as a reference point to evaluate the performance of more complex models, establishing a performance floor.

- **SeasonalNaive:** This model assumes that the forecast value is identical to the last observed value from the same season. For this monthly data, the forecast for a given month is simply the value from the same month in the previous year[24].

2.3.2. Statistical Models

Statistical models are traditional approaches that rely on well-defined mathematical assumptions about the underlying data-generating process, making them robust for series with clear trend and seasonal patterns.

- **AutoETS:** This method automatically selects the best-fitting Exponential Smoothing (ETS) state-space model by testing various configurations of error, trend, and seasonality (additive, multiplicative, damped, etc.) and choosing the one that minimizes a given information criterion[25].
- **DynamicOptimizedTheta:** An optimized implementation of the Theta decomposition model, which decomposes the time series into two "theta lines" to model short-term behavior and long-term trend separately before combining them for the final forecast.
- **NPTS (Non-Parametric Time Series):** A kernel-based forecasting method that makes no strong assumptions about the data's underlying distribution. It predicts future values by looking at past patterns, making it flexible for series that do not follow standard statistical models.

Table 2. Hyperparameters for Non-LLM Models

Model	Parameters
Baseline Model	
SeasonalNaive	• `seasonal_period = 12` • `n_jobs = 0.5`
Statistical Models	
AutoETS	• `seasonal_period = 12` • `model = "additive"`
DynamicOptimizedTheta	• `seasonal_period = 12` • `decomposition_type = "multiplicative"`
NPTS	• `kernel_type = "exponential"` • `exp_kernel_weights = 1.0` • `use_seasonal_model = True` • `num_default_time_features = 1`
Tabular Models	
RecursiveTabular	• `maxlags = 24` • `use_lags = "auto"` • `regressor = "LightGBM"`
DirectTabular	• `maxlags = 24` • `use_lags = "auto"` • `regressor = "LightGBM"`
Deep Learning Models	
DeepAR	• `context_length = 12` • `batch_size = 32` • `lr = 1e-4` • `lr_scheduler_type = "cosine"`
TemporalFusionTransformers	• `context_length = 12` • `batch_size = 32` • `lr = 1e-4` • `lr_scheduler_type = "cosine"`
TiDE	• `context_length = 12` • `batch_size = 32` • `lr = 1e-4` • `lr_scheduler_type = "cosine"`
PatchTST	• `context_length = 12` • `batch_size = 32` • `lr = 1e-4` • `lr_scheduler_type = "cosine"`

2.3.3. Tabular Models

These models transform the time series forecasting problem into a standard regression task by creating a tabular dataset where lagged historical values serve as input features to predict a future value. This allows powerful machine learning regressors, such as LightGBM, to be applied[26].

- **RecursiveTabular:** Forecasts a single step ahead and then recursively uses this new prediction as an input feature to forecast the subsequent step, repeating until the full forecast horizon is generated.
- **DirectTabular:** Trains a separate model for each step in the forecast horizon. This avoids the accumulation of errors common in the recursive strategy, as each forecast is made independently.

2.3.4. Deep Learning Models

Deep learning models utilize multi-layered neural network architectures to automatically learn complex, non-linear patterns and temporal dependencies from the data without manual feature engineering.

- **DeepAR:** An autoregressive model based on a Recurrent Neural Network (RNN) that is specifically designed for probabilistic forecasting. Instead of a single point forecast, it outputs a full probability distribution for each future time step[27].
- **PatchTST:** A Transformer-based architecture that first segments the input time series into smaller, overlapping windows or "patches." This patching mechanism allows the model's attention mechanism to more efficiently learn both local and long-range dependencies[28].
- **Temporal Fusion Transformer (TFT):** A sophisticated, attention-based architecture that fuses information from multiple sources. It uses gating mechanisms to filter irrelevant information and interprets its own temporal dynamics to produce highly accurate and interpretable forecasts[29].
- **TiDE:** An efficient forecasting model based on a simple Multi-layer Perceptron (MLP) encoder-decoder structure. It processes temporal features and covariates through dense layers, making it computationally faster than attention-based models while remaining effective for long-horizon forecasting.

2.3.5. Large Language Models (LLMs)

This category leverages pre-trained foundation models from the Chronos family, which apply transfer learning to time series forecasting. These models are pre-trained on vast amounts of public time series data and can generalize to new datasets with minimal task-specific training[30].

- **Zero-Shot Forecasting:** The pre-trained Chronos models are used directly to generate forecasts without any training on the Indonesian retail sales data. This tests their intrinsic, out-of-the-box generalization capability.
- **Fine-Tuning:** The pre-trained models are further trained on our specific dataset. This allows the model to adapt its learned patterns to the unique characteristics and dynamics of the Indonesian retail sales index.

Table 3. Hyperparameters for Fine-Tuning Chronos Models

Model	Parameters
Chronos Models	• `fine_tune = True` • `num_samples = 100` • `fine_tune_batch_size = 32` • `fine_tune_lr = 1e-4` • `lr_scheduler_type = "cosine"`

All models were trained to forecast a 12-month horizon (February 2024 through January 2025) for each category. We did not incorporate any exogenous variables during

training, to isolate the models' capability of capturing patterns from the Retail Sales Index itself.

2.4. Performance Evaluation

The performance of all forecasting models was evaluated using a comprehensive set of seven metrics, assessing both point forecast accuracy and the quality of probabilistic predictions. Lower values of all these error metrics indicate better model performance. These metrics are defined as follows:

- **Mean Absolute Error (MAE):** MAE measures the average magnitude of the forecast errors, irrespective of direction [31,32].
- **Mean Squared Error (MSE):** MSE is the average of the squared forecast errors, which penalizes larger errors more heavily [31,33].
- **Root Mean Squared Error (RMSE):** RMSE is the square root of MSE, bringing the error metric back to the original units of the data for interpretability [31,33].
- **Mean Absolute Percentage Error (MAPE):** MAPE measures the average percentage error. It is computed as the mean of the absolute percentage differences between forecasts and actuals [34].
- **Symmetric Mean Absolute Percentage Error (SMAPE):** SMAPE is a variation of MAPE that symmetrically normalizes errors [35,36].
- **Scaled Quantile Loss (SQL):** SQL evaluates the accuracy of probabilistic forecasts by measuring the quantile loss at specified quantile levels, scaled by a factor (often a normalization term for each series) [37,38].
- **Weighted Quantile Loss (WQL):** WQL is similar to SQL but assigns different weights to different quantile levels, emphasizing certain parts of the predictive distribution [37,39].

3. Result

This section presents the results of the forecasting experiments for all models described above. Performance evaluation is first summarized in tabular form, followed by graphical comparisons of metric outcomes across models, and finally a detailed examination of the fine-tuned LLM forecasts for each product category.

3.1. Evaluation Metric Comparison for non-llm Models

The performance of the non-LLM models, encompassing the baseline, statistical, tabular, and deep learning categories, is detailed in Table 4a. The evaluation reveals a distinct split in performance based on the metric type, highlighting that different models excel in either probabilistic or point forecasting tasks.

For probabilistic forecasting, the deep learning models proved superior. The Temporal Fusion Transformer achieved the best Scaled Quantile Loss (SQL) of 0.346, while TiDE delivered the lowest Weighted Quantile Loss (WQL) at 0.177. This indicates their strength in estimating accurate prediction intervals.

Conversely, for point forecast accuracy, the statistical model DynamicOptimizedTheta was the clear leader, securing the lowest Mean Absolute Error (MAE) of 0.217, Mean Squared Error (MSE) of 0.080, Root Mean Squared Error (RMSE) of 0.283, and Symmetric MAPE (SMAPE) of 0.277. This demonstrates its robustness in predicting the central tendency of the series.

In contrast, several models, including Direct Tabular, DeepAR, and especially NPTS, consistently exhibited the highest error rates across all metrics, suggesting they were less effective at capturing the complex patterns within the Indonesian retail sales data.

3.2. Evaluation Metric Results for Chronos Models (Zero-Shot)

Table 4b presents the performance metrics of the Chronos family of large language models (LLMs) when deployed in a zero-shot setting—that is, without undergoing any fine-tuning on Indonesian retail data. Notably, these pre-trained models exhibited exceptional

Table 4. Comprehensive Performance Evaluation of All Forecasting Models.

(a) Evaluation Metrics for non-LLM Performance

Model	SQL	WQL	MAE	MAPE	MSE	RMSE	SMAPE
TemporalFusionTransformer	**0.346**	0.179	0.243	0.298	0.128	0.358	0.449
TiDE	0.390	**0.177**	0.258	**0.277**	0.101	0.317	0.302
DynamicOptimizedTheta	0.418	0.203	**0.217**	0.283	**0.080**	**0.283**	**0.277**
SeasonalNaive	0.441	0.224	0.237	0.391	0.109	0.331	0.456
AutoETS	0.441	0.224	0.237	0.391	0.109	0.331	0.456
RecursiveTabular	0.441	0.238	0.260	0.422	0.101	0.318	0.285
PatchTST	0.493	0.209	0.316	0.312	0.153	0.391	0.359
DirectTabular	0.962	0.386	0.520	0.458	0.379	0.616	0.629
DeepAR	1.079	0.471	0.666	0.587	0.627	0.792	0.903
NPTS	1.582	0.677	1.126	1.019	1.449	1.204	1.749

(b) Zero-Shot Performance

Model	SQL	WQL	MAE	MAPE	MSE	RMSE	SMAPE
Chronos [small]	**0.271**	0.141	0.198	0.261	0.074	0.272	0.282
Chronos [bolt_tiny]	0.280	**0.137**	**0.178**	0.247	**0.053**	**0.230**	**0.215**
Chronos [bolt_mini]	0.282	**0.137**	0.181	0.248	0.057	0.239	0.226
Chronos [bolt_base]	0.289	0.145	0.191	0.247	0.070	0.265	0.254
Chronos [bolt_small]	0.296	0.145	0.192	**0.224**	0.067	0.259	0.246
Chronos [base]	0.309	0.146	0.203	0.258	0.071	0.267	0.246
Chronos [large]	0.328	0.151	0.220	0.268	0.076	0.275	0.269
Chronos [mini]	0.341	0.169	0.231	0.297	0.090	0.301	0.322
Chronos [tiny]	0.354	0.170	0.240	0.321	0.098	0.313	0.331

(c) Fine-Tuned Performance

Model	SQL	WQL	MAE	MAPE	MSE	RMSE	SMAPE
Chronos – FT [base]	**0.274**	**0.136**	**0.184**	**0.267**	**0.059**	**0.243**	**0.218**
Chronos – FT [small]	0.294	0.146	0.197	0.267	0.068	0.261	0.246
Chronos – FT [mini]	0.372	0.168	0.231	0.300	0.079	0.280	0.259
Chronos – FT [large]	0.374	0.183	0.245	0.339	0.111	0.333	0.270
Chronos–FT [bolt_small]	0.393	0.184	0.242	0.264	0.105	0.323	0.328
Chronos–FT [bolt_base]	0.431	0.223	0.301	0.396	0.154	0.393	0.464
Chronos–FT [bolt_mini]	0.472	0.263	0.368	0.471	0.267	0.517	0.563
Chronos–FT [bolt_tiny]	0.498	0.281	0.377	0.488	0.295	0.543	0.613
Chronos – FT [tiny]	0.592	0.270	0.381	0.422	0.216	0.465	0.418

predictive capabilities, surpassing both traditional and deep learning approaches across nearly all evaluation criteria.

Among them, the lightweight Chronos [bolt_tiny] model stood out as the best-performing variant, recording the lowest values for SQL (0.280), WQL (0.137), MAE (0.178), MSE (0.053), RMSE (0.230), and SMAPE (0.215). Other compact models, including Chronos [bolt_mini] and Chronos [small], also demonstrated highly competitive performance.

These zero-shot outcomes underscore the substantial promise of foundation models in time-series forecasting tasks. The ability of Chronos to generalize temporal patterns learned from large-scale public corpora to the Indonesian retail sales domain—without any exposure to the target data—reflects its robust generalization capacity and potential for real-world deployment.

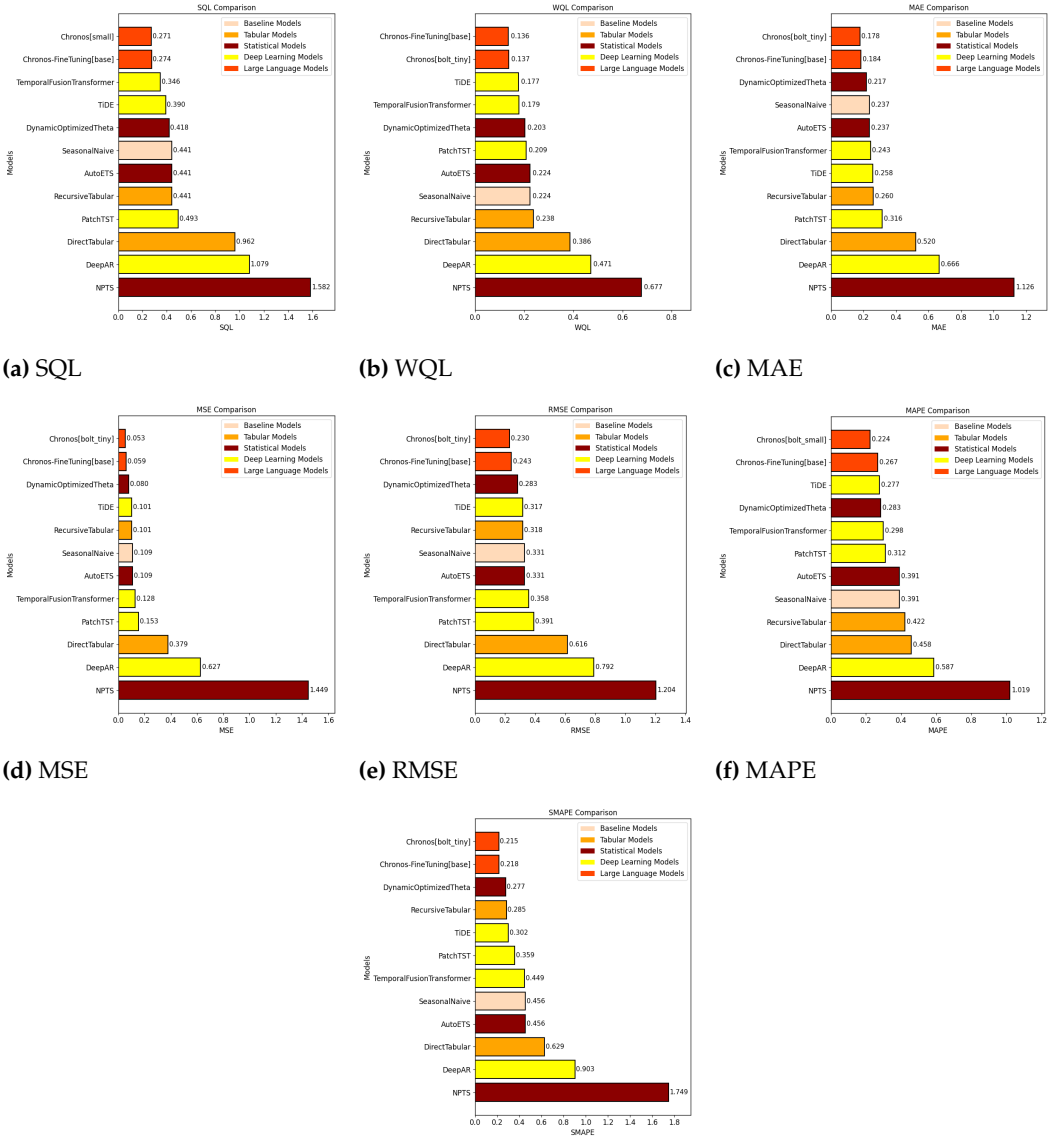
3.3. Evaluation Metric Results for Chronos Models (Fine-Tuned)

Table 4c presents the performance of the Chronos LLM models after being fine-tuned on the Indonesian retail sales training data.

The fine-tuning process further solidified the advantage of LLMs, with the Chronos – FT [base] model achieving the best performance overall. It recorded the lowest values across the majority of metrics, with an SQL of 0.274, WQL of 0.136, MAE of 0.184, MAPE of 0.267, MSE of 0.059, RMSE of 0.243, and SMAPE of 0.218. These results represent a significant improvement over both the zero-shot LLMs and all traditional models evaluated.

An interesting pattern emerged regarding model size. Fine-tuning provided a clear benefit for the standard Chronos models (base, small, mini, large), allowing them to adapt more closely to the dataset's specific characteristics. However, for the more compact bolt variants, fine-tuning led to a notable degradation in performance compared to their zero-shot counterparts. For instance, the SQL for Chronos–FT [bolt_tiny] (0.498) was substantially worse than its zero-shot version (0.280). This suggests that the smaller bolt models may be prone to overfitting when fine-tuned on a dataset of this size, and their strength lies in their powerful zero-shot capabilities.

3.4. Comparative Performance Analysis



(a) SQL **(b)** WQL **(c)** MAE

(d) MSE **(e)** RMSE **(f)** MAPE

(g) SMAPE

Figure 1. Comparison of Forecasting Models Across All Evaluation Metrics

For a quantitative and visual assessment of model efficacy, a comparative analysis was conducted using the graphical representations depicted in Figure 1. Each bar chart corresponds to a specific evaluation metric, where a lower magnitude indicates superior model performance. The models are systematically color-coded by their architectural category: Baseline (cream), Tabular (orange), Statistical (maroon), Deep Learning (yellow), and LLM-based (red). A primary observation across all evaluated metrics is the distinct

performance stratification, wherein the LLM-based Chronos models consistently demonstrate a higher degree of accuracy compared to their traditional and other deep learning counterparts.

3.4.1. Evaluation of Probabilistic Forecasting Accuracy (SQL & WQL)

Given that the principal objective of this research is probabilistic forecasting, the Scaled Quantile Loss (SQL) and Weighted Quantile Loss (WQL) serve as the most critical performance indicators. The analysis of Figures 1(a) and 1(b) reveals the unambiguous superiority of the Chronos models in terms of probabilistic calibration.

- **SQL Analysis:** The fine-tuned 'Chronos – FT [base]' (SQL = 0.274) and zero-shot 'Chronos [small]' (SQL = 0.271) exhibit nearly indistinguishable top-tier performance. These models significantly outperform the best non-LLM competitor, 'TemporalFusionTransformer' (SQL = 0.346), indicating a substantial improvement in the accuracy of the predicted quantiles.
- **WQL Analysis:** In the context of weighted quantile loss, the 'Chronos-FT [base]' model emerges as the definitive model with the best performance with a WQL of 0.136. Its performance is marginally, but consistently, better than the closest zero-shot competitors, 'Chronos [bolt_tiny]' and 'Chronos [bolt_mini]' (WQL = 0.137). This result highlights the model's enhanced ability to accurately estimate the entire predictive distribution with appropriate weighting.

3.4.2. Evaluation of Point Forecast Accuracy (MAE, MSE & RMSE)

The metrics of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), as illustrated in Figures 1(c), 1(d), and 1(e), offer a quantitative perspective on the central tendency accuracy of the forecasted values. Notably, both MSE and RMSE incorporate a squared loss component, thereby disproportionately penalizing larger prediction errors.

Once again, the superiority of the Chronos architecture is evident. The zero-shot variant Chronos [bolt_tiny] achieves the lowest errors across all three metrics (MAE = 0.178, MSE = 0.053, RMSE = 0.230). Meanwhile, the fine-tuned model Chronos – FT [base] delivers comparably strong results (MAE = 0.184, MSE = 0.059, RMSE = 0.243). The relatively low MSE and RMSE values obtained by these LLMs reflect their resilience and reduced likelihood of producing extreme forecasting errors—a notable shortcoming observed in models such as NPTS and DeepAR, which yield significantly higher error magnitudes.

3.4.3. Evaluation of Proportional Accuracy (MAPE & SMAPE)

The scale-independent metrics Mean Absolute Percentage Error (MAPE) and Symmetric Mean Absolute Percentage Error (SMAPE) evaluate forecast accuracy in proportion to the magnitude of actual observations. As shown in Figures 1(f) and 1(g), models based on large language models (LLMs) continue to exhibit superior performance.

The fine-tuned Chronos – FT [base] model achieves the best SMAPE score of 0.218, reflecting the most consistent proportional accuracy among all evaluated approaches. Although the lowest MAPE is recorded by a zero-shot variant, the consistently strong results of the fine-tuned base model across both metrics highlight its capacity to produce forecasts that are not only precise in absolute terms but also reliably scaled across diverse time series.

3.4.4. Synthesis of Comparative Findings

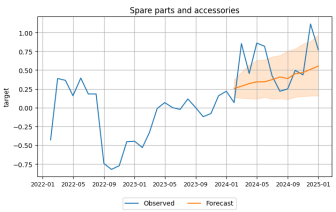
The comprehensive evaluation consistently indicates the superior predictive power of the Chronos-based LLMs. While optimal performance on specific point-forecast metrics is occasionally achieved by a zero-shot variant, the fine-tuned 'Chronos [base]' model presents the most holistically robust and effective performance profile.

Its preeminence is established by its top-tier results in the primary probabilistic evaluation criteria (SQL and WQL), aligning directly with the research's main objective. This,

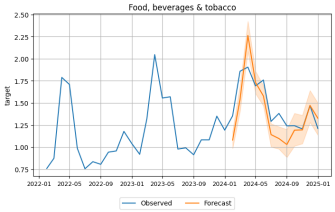
combined with its highly competitive standing across all other point and proportional error metrics, substantiates the conclusion that 'Chronos – FT [base]' is the most efficacious and reliable forecasting model for the Indonesian Retail Sales Index dataset investigated in this study.

3.5. Forecast Visualization and Category-wise Analysis

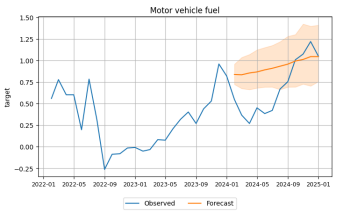
To gain deeper insights, we plotted the forecast results of the **Chronos – Fine Tuning [base]** model against the actual retail index for each of the seven product categories (Figure 2). In each chart, the blue line denotes the actual observed sales index, and the orange line denotes the Chronos model's forecast for February 2024 to January 2025. The shaded orange band represents the model's predicted confidence interval (derived from the quantile forecasts), reflecting the uncertainty in the predictions. The following subsections discuss each category in detail.



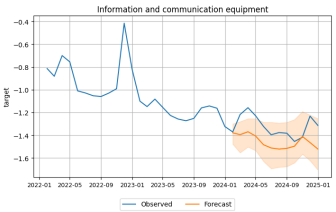
(a) Spare Parts and Accessories



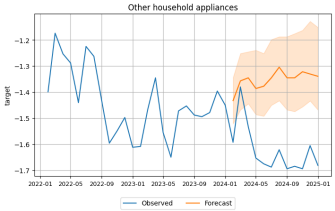
(b) Food, Beverages & Tobacco



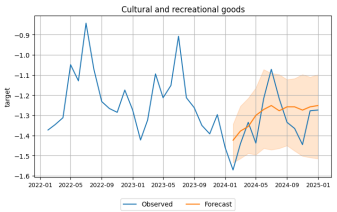
(c) Motor Vehicle Fuel



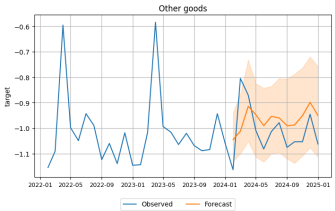
(d) Information & Comm Equipment



(e) Other Household Appliances



(f) Cultural and Recreational Goods



(g) Other Goods

Figure 2. Forecast vs. Actuals for Each Retail Category using Fine-Tuned Chronos [base]

3.5.1. Spare Parts and Accessories

Figure 2(a) shows the forecast for the "Spare Parts and Accessories" category. The actual sales in this category exhibit a somewhat volatile pattern with an underlying mild upward trend toward the end of the forecast period. The Chronos model successfully captures the general upward movement (as evidenced by a relatively low RMSE of 0.243 for this series) but tends to smooth out the short-term volatility. Notably, it misses some sharp fluctuations, for instance the peaks in late 2024 and in January 2025. The forecast often overpredicts during mid-2024 and slightly under-predicts in the final months, indicating the model's tendency to average out the noise. The confidence interval widens over the forecast horizon, reflecting increasing uncertainty further out in time, yet it remains reasonably calibrated around the actual values. In fact, the model's excellent probabilistic calibration is indicated by its low aggregate SQL (0.274). Overall, for Spare Parts and Accessories,

Chronos [base] provides a solid grasp of trend but underestimates sudden jumps, which could imply caution for inventory decisions that need to account for such irregular spikes.

3.5.2. Food, Beverages & Tobacco

Figure 2(b) shows the forecast for the "Food, Beverages & Tobacco" category. The actual data for this category have a pronounced spike in early 2024 followed by a decline and then considerable volatility. This early-2024 spike is likely tied to a seasonal event (e.g., increased spending around Ramadan, which is a major festive period in Indonesia)[40]. The Chronos model did not anticipate this sharp surge; instead, it forecasted a much smoother trajectory through that period. Consequently, its error during the spike is significant (reflected in a relatively higher MAPE of 0.267 for this category). After the spike, the model does capture the general downward trend, but it underestimates the magnitude of subsequent fluctuations in late 2024 and January 2025. The confidence interval balloons around the time of the spike—indicating the model's uncertainty in that volatile period—and then narrows post-spike as the volatility decreases. The model's Weighted Quantile Loss for this category (WQL = 0.136) is low, suggesting that despite missing the exact spike, it did provide a reasonable probabilistic envelope. However, from an operational standpoint, failing to predict the Ramadan-related surge could lead retailers to understock during peak demand, highlighting a scenario where domain-specific knowledge (like incorporating holiday effects) could improve forecasts.

3.5.3. Motor Vehicle Fuel

Figure 2(c) shows the forecast for the "Motor Vehicle Fuel" category. The actual sales for motor fuel display a clear upward trend over time with periodic dips (likely owing to seasonal patterns such as holidays or fluctuations in fuel prices)[41]. The Chronos model performs well in tracking the overall upward trend, which is evidenced by a low MAE of 0.184 for this series. However, it tends to smooth over the smaller seasonal dips. For instance, dips observed in early 2024 and around the new year of 2025 are not fully reflected in the forecast, causing the model to under-predict demand during those specific periods. Toward the end of the forecast horizon, the model slightly over-predicts the actual values. The confidence interval for the fuel category remains relatively narrow compared to more volatile categories, indicating high confidence in the trend direction. This suggests that the model is quite certain about the steady growth trajectory. From a planning perspective, the Chronos forecast for motor fuel seems reliable for long-term decisions (e.g., anticipating continuous growth in fuel demand), but it may not fully capture short-term downward deviations (e.g., temporary drops in fuel sales due to travel restrictions or unusual weather), which could lead to oversupply in those short lulls if not adjusted manually.

3.5.4. Information and Communication Equipment

Figure 2(d) shows the forecast for the "Information and Communication Equipment" category. The actual series for information and communication equipment exhibits a declining long-term trend with intermittent spikes (for example, noticeable bumps in mid-2022 and early 2023, which might correspond to new product launches or major promotional events driving temporary sales increases)[42]. The Chronos model predicts a continued gradual decline over the forecast period, which aligns with the general downward trajectory of the recent actuals. In fact, the forecast's smooth decline fits the trend but fails to capture the volatility around late 2024 and January 2025, instead smoothing through those variations. This can be seen in the model underestimating the magnitude of short-term deviations—resulting in forecasts that are sometimes above or below the realized values when the actual series briefly deviates from the trend. The SMAPE for this category is 0.218, indicating the model's percentage errors are moderate. The confidence interval widens slightly over time, reflecting growing uncertainty further out, but remains contained, indicating the model's reasonable confidence in the downward trend projection. The tendency to overlook short-term jumps could limit the model's usefulness for anticipating sudden

surges in demand for this category (e.g., a spike due to a new smartphone release). Retailers might need to supplement this with external information about upcoming product launches or promotions to avoid understocking during those events.

3.5.5. Other Household Appliances

Figure 2(e) shows the forecast for the "Other Household Appliances" category. The actual sales for other household appliances are highly volatile, with sharp peaks and troughs throughout the historical period. Despite this volatility, there appears to be a mild upward drift over the long term. The Chronos model's forecast in this category presents a relatively steady, gently rising trajectory. This smoothed forecast fails to reproduce the wild swings present in the actual data. For example, a deep trough in mid-2024 followed by a rapid recovery is largely missed by the model, which predicts a much higher value than the trough (over-predicting the low point) and a lower value than the actual rebound (under-predicting the subsequent peak). This is indicative of the model effectively averaging out the extremes. As a result, during a low-demand period the model overshoots (potentially suggesting stocking too much inventory), and during a high-demand rebound it undershoots (risking stockouts). The MSE for this category is relatively low at 0.059, which might be somewhat misleading because the model never deviates hugely due to its smoothing, but it consistently does not match the amplitude of changes. The forecast's confidence interval is wide, reflecting the model's awareness of uncertainty in this erratic series. The persistent wide interval around the forecast underscores the difficulty of this category: the model recognizes its limitations here. This category's behavior suggests that purely endogenous forecasting (using only past sales) may be insufficient; incorporating exogenous information such as promotional calendars or macroeconomic indicators could be necessary to anticipate the large swings. In summary, while the Chronos model captures the general trend for household appliances, its smoothing of extreme fluctuations could reduce its immediate utility for operational planning unless combined with volatility-targeted adjustments.

3.5.6. Cultural and Recreational Goods

Figure 2(f) shows the forecast for the "Cultural and Recreational Goods" category. The actual sales in the cultural and recreational goods category are extremely volatile, with significant spikes (e.g., large surges in 2022 and 2023) and subsequent drops. In the forecast period (2024–2025), the actual trend appears to waver without a clear direction, though the model predicts a slight upward drift. The Chronos model's forecast smooths over these fluctuations, resulting in a mild increase that fails to capture the dramatic peak in mid-2024 or the drop that follows. During the big spike in mid-2024, the model underestimates demand (because it never predicted such a spike), and after the spike, it overestimates when the actual sales fall off. This pattern of under-predicting peaks and over-predicting troughs is a common theme for the model in highly volatile series. Consequently, the SQL for this category (0.274, aggregated) is higher than in more stable categories, indicating the model's quantile estimates had larger errors relative to actuals. The confidence interval for the forecast is very broad around the time of the spike, which correctly signals high uncertainty but does not correct the mean forecast itself. The sheer unpredictability of this category means that even a sophisticated model like Chronos struggles to provide accurate point forecasts. Retailers dealing in cultural and recreational goods would likely need to hedge against this uncertainty, perhaps by maintaining flexible inventory or by incorporating real-time demand signals. The model's performance here underscores a limitation: LLMs excel at learning patterns, but when the data is dominated by one-off events or irregular spikes (which might be driven by external factors such as one-time events or fads), the model's inherently smooth prediction fails to capture those irregularities.

3.5.7. Other Goods

Figure 2(g) shows the forecast for the "Other Goods" category. The "Other Goods" category is a catch-all that also exhibits high volatility. The historical data show notable spikes in 2022 and 2023, followed by a decline in late 2023 and renewed fluctuations during 2024. Specifically, mid-2024 features a sharp peak (perhaps due to a major event or promotion) and then a drop with oscillations thereafter. The Chronos model forecasts a fairly steady upward trend through 2024 into early 2025, largely missing the mid-2024 peak and the subsequent volatility. During the actual mid-2024 surge, the model's forecast is significantly lower (under-predicting the true demand), and following that, when actual sales dip, the model's forecast is higher (over-predicting). This again is the smoothing effect we observed in other volatile categories. The WQL for this category is 0.136, which, interestingly, is relatively low—implying the model's prediction intervals might have been wide enough to partially cover the variability, even if the mean line was off. Indeed, the forecast shows a very wide confidence interval around the mid-2024 period, indicating the model's uncertainty was greatest where the actual variance was high. Practically speaking, the wide confidence bands mean that while the point forecast is smoothed, a risk-aware retailer could use the prediction interval to prepare for best- and worst-case scenarios. Nonetheless, the utility of the mean forecast for short-term planning is limited. For a product mix as diverse as "Other Goods," unmodeled factors (like seasonal events or irregular demand surges for specific items in this category) appear to drive the unpredictability. This suggests that future improvements could include segmenting this category further or adding external inputs to help predict those irregular changes.

3.5.8. General Insights Across Categories

Considering all categories together, the fine-tuned Chronos [base] model demonstrates an impressive ability to capture long-term trends and overall direction of change. This is reflected in the consistently low overall error metrics (e.g., a global RMSE of 0.243 and MAE of 0.184). However, a clear limitation is its tendency to smooth out abrupt short-term fluctuations. This was most evident in categories with high volatility, such as "Other Household Appliances," "Cultural and Recreational Goods," and "Other Goods," where the model struggled to anticipate sharp peaks and troughs. This behavior is likely attributable to the transformer-based architecture of the LLM, which emphasizes learning global patterns and might de-emphasize isolated local anomalies unless they are recurrent.

On a positive note, the probabilistic nature of the Chronos forecasts is a strength. The model's confidence intervals generally widened appropriately during volatile periods and further into the forecast horizon. This suggests that while the point forecasts may miss extreme values, the model still conveys useful information about uncertainty through SQL/WQL metrics and interval spread. In a practical sense, these probabilistic forecasts can be valuable for risk management—retailers can gauge the range of potential outcomes and prepare contingency plans for high-demand or low-demand scenarios.

In summary, the Chronos [base] LLM excels in capturing broad trends in Indonesia's retail sales data, indicating its suitability for strategic forecasting and long-term planning. For short-term, volatility-sensitive applications, there may be benefit in augmenting the LLM approach with techniques or data that specifically target those local variations (for example, a hybrid model or inclusion of external signals related to promotions or holidays).

4. Discussion

4.1. Model Performance and Architectural Insights

The empirical findings clearly demonstrate the superior performance of the fine-tuned Chronos [base] model, which achieved the most favorable outcomes across key evaluation metrics (SQL = 0.274, WQL = 0.136, MAE = 0.184, SMAPE = 0.218). These results underscore the significant potential of transformer-based large language models (LLMs) for probabilistic time series forecasting, aligning with growing evidence supporting the effectiveness of large sequence models in capturing complex temporal structures [15]. The

strong performance of Chronos [base] can be attributed to two primary factors: its underlying transformer architecture—well-suited for modeling long-range dependencies—and the fine-tuning procedure, which enabled the model to adapt its general pre-trained knowledge to the specific statistical characteristics of Indonesian retail sales data.

An intriguing pattern emerged when comparing the performance of the standard Chronos models and their more compact bolt variants after fine-tuning. While fine-tuning improved the performance of larger models, it unexpectedly led to a decline in the effectiveness of the smaller bolt models, which had performed exceptionally well in the zero-shot setting. This points to a potential trade-off: compact models may harbor a highly generalized and robust pre-trained state that becomes vulnerable to overfitting when exposed to limited, domain-specific data, whereas larger models have the capacity to benefit more from fine-tuning due to their greater representational flexibility. Additionally, the tendency of Chronos models to smooth out short-term volatility in forecasts appears to reflect an architectural bias inherent to transformers. By design, transformers aggregate information across entire sequences, which facilitates the extraction of global patterns but may inadvertently treat abrupt, localized variations as noise [43]. While this quality supports accurate trend detection, it may hinder the model's responsiveness to sudden, meaningful shifts in market behavior.

4.2. Comparison with Existing Literature

The results of this study position Chronos-based large language models (LLMs) as a substantial advancement over many conventional forecasting approaches. Although traditional statistical methods such as AutoETS and DynamicOptimizedTheta exhibited commendable performance, they were consistently outperformed by the fine-tuned Chronos [base] model particularly in terms of probabilistic evaluation metrics like SQL and WQL. This superiority underscores not only the enhanced point forecast accuracy of LLMs but also their improved ability to quantify uncertainty, a critical component in contemporary risk assessment and inventory planning.

When compared to other deep learning architectures, including the advanced Temporal Fusion Transformer (TFT), Chronos models retained a clear performance advantage. For example, the most accurate non-LLM model on the SQL metric—TFT—achieved a score of 0.346, markedly higher than the 0.274 recorded by Chronos [base]. This discrepancy highlights the efficacy of the extensive pre-training undertaken by Chronos, which appears to endow it with the capability to recognize nuanced patterns within a moderately sized dataset more effectively than models trained exclusively on the task-specific data. These findings align with a broader trend in artificial intelligence, whereby large, pre-trained foundation models demonstrate superior generalization and data efficiency [13]. While prior research has shown promise in hybrid architectures [8], the present results indicate that a well-optimized LLM can independently manage considerable modeling complexity. Nonetheless, integrating LLMs with domain-specific components remains a promising avenue for addressing localized volatility patterns that Chronos tends to smooth over.

4.3. Implications for the Indonesian Retail Sector

The results of this study have several implications for retail businesses and policymakers in Indonesia. First, the ability of Chronos [base] to accurately capture long-term trends (e.g., its strong performance in categories like Motor Vehicle Fuel with low MAE of 0.184) means that retailers can rely on such models for strategic planning. For example, a consistent upward trend predicted in fuel sales can inform capacity expansion, procurement, and supply chain adjustments well in advance. Similarly, in categories where the model predicts a decline or plateau, businesses can proactively manage inventory levels to avoid overstocking.

The probabilistic nature of the forecasts (low SQL and WQL scores) is particularly valuable in a volatile economic environment. Retail sales can be influenced by sudden policy changes, economic shocks, or shifts in consumer [44]. The confidence intervals

provided by the LLM forecasts allow decision-makers to gauge the uncertainty and plan for best-case and worst-case scenarios. For instance, a retailer seeing a wide interval for an upcoming festive season might decide to secure extra stock as a buffer, whereas a narrow interval might instill confidence in a more aggressive lean inventory approach.

On the other hand, the model's tendency to smooth short-term fluctuations signals a caution. In categories like Cultural and Recreational Goods, where demand spikes can be large and sudden, a sole reliance on the LLM forecast could result in missed opportunities or losses (e.g., running out of stock when a spike hits, or tying up capital in inventory that isn't immediately needed during a slump). Thus, for operational short-term forecasting (say, week-to-week stock ordering for a promotion), businesses may need to complement the LLM's forecast with local knowledge or short-term adjustment factors. An approach could be to use the LLM's forecast as a baseline and then overlay known upcoming events (holiday sales, marketing campaigns) to adjust the predictions.

For policymakers, accurate retail forecasts can support macroeconomic planning. The retail sales index is often a bellwether for consumer confidence and economic health. Improved forecasting using LLMs can enhance the ability of institutions like Bank Indonesia to anticipate economic turning points or to gauge the impact of events (for example, how quickly retail spending might recover after a downturn). The model's success here suggests that similar approaches could be extended to other economic indicators for a more data-driven policy planning process.

4.4. Limitations

This study is subject to several limitations. First, the primary limitation is the architectural tendency of the Chronos models to smooth over high-frequency volatility and fail to predict abrupt, one-off demand spikes. This is likely a consequence of the model's global attention mechanism and the exclusive use of endogenous data, without the context of external factors that often drive such events. Second, our analysis was constrained to seven aggregated national-level product categories. The model's performance and generalizability on more granular data, such as SKU- or store-level sales, which exhibit different statistical properties (e.g., intermittency), remain to be validated. Third, the study did not incorporate exogenous variables (e.g., holidays, economic indicators), the inclusion of which could potentially mitigate the issue of unpredictable demand spikes. Lastly, the research was conducted in a batch forecasting setting; the model's adaptability and performance in a real-time, continuously updating environment have not been assessed.

4.5. Future Research Directions

The encouraging results of leveraging LLMs for retail forecasting open up several avenues for future work:

1) **Hybrid Modeling:** As suggested, a hybrid approach could marry the strengths of LLMs with those of traditional models. For instance, one could use a fine-tuned Chronos model to forecast the general trend and then apply a lightweight statistical model (or even rule-based adjustments) to account for known upcoming irregular events. Such a combination might address the short-term fluctuation issue. Research could explore architectures where an LLM provides one component of the forecast and a classical model adjusts the residuals (or vice versa), akin to how ensembles are built to capture both linear and non-linear components.

2) **Incorporating Exogenous Variables:** The current study was univariate for each category (using only the sales history). Many factors influence retail sales: economic indicators (inflation, interest rates), calendar events (holidays like Eid, Christmas, Chinese New Year, etc.), and even weather or social trends. Future research could integrate these exogenous variables into the forecasting process. Large models like LLMs can, in principle, handle multiple inputs; an interesting direction would be to fine-tune or prompt LLMs with auxiliary information (for example, feeding a sequence of indicator values along with sales data to forecast future sales). Preliminary steps could include encoding holiday information

or incorporating Google Trends data for product categories to see if that reduces error around spikes.

3) **Addressing Metric Anomalies:** The issue of negative error metrics in our AutoGluon evaluation needs resolution. Future work should replicate the evaluation with an independent computation of all metrics (perhaps by exporting forecasts and calculating metrics in a controlled environment). Additionally, exploring alternative probabilistic metrics like the Continuous Ranked Probability Score (CRPS) could provide a more holistic assessment of forecast distributions, as CRPS is a single-number metric that generalizes quantile metrics and is widely used in probabilistic forecasting competitions. Using such metrics might also make our results more comparable with other studies.

4) **Real-time and Adaptive Forecasting:** The retail environment can change rapidly (as witnessed during the COVID-19 pandemic). A worthwhile direction is to test the LLM forecasting approach in a real-time setting, where the model is periodically updated with new data and has to adapt to potential structural changes in the series. Investigating the robustness of an LLM like Chronos when the underlying demand pattern shifts (say, due to an economic shock) would be valuable. Does the pre-trained knowledge help it adapt faster, or would it also suffer until retrained? Perhaps techniques like online learning or transfer learning could be employed to keep the model in sync with the latest trends without full retraining.

5) **Granular Forecasting and Hierarchies:** Our use case was at a fairly aggregated level (national index for broad categories). Future research could apply LLMs to more granular data, such as individual store sales or specific product lines, and also explore hierarchical forecasting (ensuring consistency between, say, store-level forecasts and total category forecasts). LLMs could be trained in a multi-series context to borrow strength across series, which AutoGluon does implicitly by treating the seven categories together for modeling. Evaluating how well the LLM approach scales to hundreds of related time series (e.g., all SKUs in a store) could be impactful for large retailers.

In conclusion, while this study demonstrates the promise of LLMs like Chronos for retail forecasting, it also highlights the need for careful integration with domain knowledge and other methods. By addressing the above directions, future work can build on our findings to create forecasting systems that are both highly accurate and practically robust, thereby providing even greater value to the retail industry.

5. Conclusion

In this study, we conducted a comprehensive evaluation of various forecasting models, culminating in the application of Chronos-based Large Language Models for the probabilistic forecasting of the Indonesian Retail Sales Index. Our findings conclusively demonstrate that a fine-tuned Chronos [base] model provides a state-of-the-art solution for this task, establishing its superiority over traditional statistical, tabular, and other deep learning models. The model achieved an outstanding overall performance, registering top-tier results in the critical probabilistic metrics of SQL (0.274) and WQL (0.136), while also maintaining highly competitive accuracy across point forecast metrics, including an MAE of 0.184 and an SMAPE of 0.218.

The research highlights the transformative potential of pre-trained foundation models in the time series domain. These models excel at identifying and projecting long-term trends and providing reliable uncertainty estimates, which are invaluable for strategic business planning and risk management. However, we also identified a key limitation: a consistent tendency to smooth over short-term, high-amplitude fluctuations, potentially leading to inaccuracies in short-horizon operational forecasting.

For practical application in the Indonesian retail sector, we recommend a hybrid approach where the LLM serves as a robust baseline for long-term strategy, complemented by other methods or domain knowledge to account for short-term volatility. Future research should focus on enhancing these models by incorporating exogenous variables, exploring hybrid architectures, and validating their performance at a more granular level.

In summary, while acknowledging current limitations, this work confirms that leveraging large language models represents a significant step forward, enabling more accurate, reliable, and data-driven decision-making in the complex and dynamic retail market.

References

1. Zhang, W. Implementation of Bigdata Techniques in Sales Forecasting: Evidence from Retailing and E-commerce. *Advances in Economics, Management and Political Sciences* **2024**, *128*. https://doi.org/10.54254/2754-1169/2024.18262.
2. Sharma, H.; et al. Enhancement of Sales Forecasting and Prediction with Machine Learning Methods. *International research journal of computer science* **2024**, *11*. https://doi.org/10.26562/irjcs.2024.v1111.02.
3. Gao, Y. A Comparative Study of ARIMA and ETS Models for Time Series Forecasting. *Advances in Economics, Management and Political Sciences* **2025**, *149*. https://doi.org/10.54254/2754-1169/2024.19252.
4. Teoh, T.T.; Cho, S.Y.; Nguwi, Y.Y. Emotional prediction using time series multiple-regression genetic algorithm for autistic syndrome disorder. In Proceedings of the 2012 7th International Conference on Computer Science Education (ICCSE), 2012, pp. 9–12. https://doi.org/10.1109/ICCSE.2012.6295015.
5. Khashei, M.; Bijari, M.; Ardali, G.A.R. Improvement of Auto-Regressive Integrated Moving Average models using Fuzzy logic and Artificial Neural Networks (ANNs). *Neurocomputing* **2009**, *72*, 956–967.
6. Ye, J.; Yang, L. A comparative study of ensemble support vector regression methods for short-term load forecasting. In Proceedings of the 2018 5th International Conference on Systems and Informatics (ICSAI). IEEE, 2018, pp. 139–143.
7. Tang, T.; Liu, T.; Gui, G. Forecasting precipitation and temperature evolution patterns under Climate Change using a Random Forest Approach with Seasonal Bias correction. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2024**.
8. Hossain, M.S.; et al. Enhanced market trend forecasting using machine learning models: a study with external factor integration. *Journal of global education and research* **2025**, *06*. https://doi.org/10.55640/business/volume06issue01-02.
9. Ni, L.; Li, Y.; Wang, X.; Zhang, J.; Yu, J.; Qi, C. Forecasting of forex time series data based on deep learning. *Procedia computer science* **2019**, *147*, 647–652.
10. Nasution, A.H.; Onan, A. ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access* **2024**, *12*, 71876–71900. https://doi.org/10.1109/ACCESS.2024.3402809.
11. Khalila, Z.; Nasution, A.H.; Monika, W.; Onan, A.; Murakami, Y.; Radi, Y.B.I.; Osmani, N.M. Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. *International Journal of Advanced Computer Science and Applications* **2025**, *16*. https://doi.org/10.14569/IJACSA.2025.01602134.
12. Nasution, A.H.; Monika, W.; Onan, A.; Murakami, Y. Benchmarking 21 Open-Source Large Language Models for Phishing Link Detection with Prompt Engineering. *Information* **2025**, *16*. https://doi.org/10.3390/info16050366.
13. Brown, T.B.; et al. Language models are few-shot learners. In Proceedings of the Advances in Neural Information Processing Systems, 2020, Vol. 33, pp. 1877–1901.
14. Hidayat, F.; Nasution, A.H.; Ambia, F.; Putra, D.F.; Mulyandri. Leveraging Large Language Models for Discrepancy Value Prediction in Custody Transfer Systems: A Comparative Analysis of Probabilistic and Point Forecasting Approaches. *IEEE Access* **2025**, *13*, 65643–65658. https://doi.org/10.1109/ACCESS.2025.3560254.
15. Zhou, Q.; et al. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning, 2021, Vol. 139.
16. Khairurrahman, A.L.; Badriah, L.S.; Sambodo, H.; Rahajuni, D.; Kadarwati, N.; Pudjianto, H.; Setiarso, O.; Supriadi, D.; Priyono, R. Informative Industrial Analytic for Effective Retail Business Performance: A Case of Emerging Economy. *WSEAS Transactions on Systems* **2023**, *22*, 170–179.
17. Cheng, K.M.; Wijaya, L.; Ng, K.C.; Anthonysamy, L. Decoding Consumer Behaviour in Indonesian E Commerce: A Stimulus-Organism-Response Analysis. *Australian Journal of Telecommunications and the Digital Economy* **2024**, *12*. https://doi.org/10.18080/jtde.v12n4.1009.
18. Suresh, B.; Suresh, M. A comprehensive analysis of retail sales forecasting using machine learning and deep learning methods. In Proceedings of the 2023 International Conference on Data Science and Network Security (ICDSNS). IEEE, 2023, pp. 1–5.
19. Ganguly, P.; Mukherjee, I. Enhancing Retail Sales Forecasting with Optimized Machine Learning Models. In Proceedings of the 2024 4th International Conference on Sustainable Expert Systems (ICSES). IEEE, 2024, pp. 884–889.
20. Bank Indonesia. Laporan Survei Penjualan Eceran (SPE) Januari 2025, 2025. [Online]. Available: https://www.bi.go.id/id/publikasi/laporan/Pages/SPE-Januari-2025.aspx. [Accessed Mar. 11, 2025].
21. Pranolo, A.; et al. Enhanced Multivariate Time Series Analysis Using LSTM: A Comparative Study of Min-Max and Z-Score Normalization Techniques. *Ilkom Jurnal Ilmiah* **2024**, *16*. https://doi.org/10.33096/ilkom.v16i2.2333.210-220.
22. Jantima, P.; Bancha, L.; Khanista, N. Comparative Study for Data Normalization Methods on Predicting Cryptocurrency Price. *-B:* **2022**, *13*, 853.
23. Phan, Q.T.; Wu, Y.K.; Phan, Q.D. An Overview of Data Preprocessing for Short-Term Wind Power Forecasting. In Proceedings of the 2021 7th International Conference on Applied System Innovation, ICASI 2021, 2021. https://doi.org/10.1109/ICASI52993.2021.9568453.

24. Caillault, É.P.; Bigand, A.; et al. Comparative study on univariate forecasting methods for meteorological time series. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018, pp. 2380–2384.
25. Pereira, P.J.; Costa, N.; Barros, M.; Cortez, P.; Durães, D.; Silva, A.; Machado, J. A comparison of automated time series forecasting tools for smart cities. In Proceedings of the EPIA Conference on Artificial Intelligence. Springer, 2022, pp. 551–562.
26. Bahrpeyma, F.; Roantree, M.; McCarren, A. Multi-resolution forecast aggregation for time series in agri datasets **2017**.
27. Amjad, F.; Korotko, T.; Rosin, A. Forecasting PV Energy Generation Using Transformer-Based Architectures: A Comparative Study of Lag-Llama, TFT, and DeepAR. In Proceedings of the 2024 IEEE 65th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON). IEEE, 2024, pp. 1–6.
28. Oliveira, J.M.; Ramos, P. Evaluating the Effectiveness of Time Series Transformers for Demand Forecasting in Retail. *Mathematics* **2024**, *12*, 2728.
29. Lim, B.; Arık, S.Ö.; Loeff, N.; Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* **2021**, *37*, 1748–1764.
30. Ansari, A.F.; Stella, L.; Turkmen, C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S.S.; Arango, S.P.; Kapoor, S.; et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815* **2024**.
31. Vishwanath, A.; Basheeruddin, M.; Saveetha, D. Forecasting Sales Data Using Time Series Models and LSTM Model. In Proceedings of the 2024 7th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), jul 2024. https://doi.org/10.1109/icait61638.2024.10690408.
32. Tello, A.; Izquierdo, I.; Pacheco, G.; Vanegas, P. Prediction of imports of household appliances in ecuador using LSTM networks. In Proceedings of the Information and Communication Technologies of Ecuador (TIC. EC). Springer, 2020, pp. 194–207.
33. Sonata, I.; Heryadi, Y. Comparison of LSTM and Transformer for Time Series Data Forecasting. In Proceedings of the 2024 7th International Conference on Informatics and Computational Sciences (ICICoS). IEEE, 2024, pp. 491–495.
34. Schwartz, Z.; Ma, J.; Webb, T. The MSapeMER: a symmetric, scale-free and intuitive forecasting error measure for hospitality revenue management. *International Journal of Contemporary Hospitality Management* **2023**. https://doi.org/10.1108/ijchm-01-2023-0088.
35. Castillo Estrada, M.d.R.; Gómez Camarillo, M.E.; Sánchez Parraguirre, M.E.; Gómez Castillo, M.E.; Meneses Juárez, E.; Cruz Gómez, M.J. Evaluation of Several Error Measures Applied to the Sales Forecast System of Chemicals Supply Enterprises. *International Journal of Business Administration* **2020**, *11*. https://doi.org/10.5430/IJBA.V11N4P39.
36. Mathai, A.V.; Agarwal, A.; Angampalli, V.; Narayanan, S.; Dhakshayani, E. Development of new methods for measuring forecast error. *International Journal of Logistics Systems and Management* **2016**, *24*, 213–225.
37. Kovalevsky, V.E.; Zhukova, N.A. Building a Model for Time Series Forecasting using AutoML Methods. In Proceedings of the 2024 XXVII International Conference on Soft Computing and Measurements (SCM). IEEE, 2024.
38. Cui, W.; Wan, C.; Song, Y. Ensemble deep learning-based non-crossing quantile regression for nonparametric probabilistic forecasting of wind power generation. *IEEE Transactions on Power Systems* **2022**, *38*, 3163–3178.
39. Lopez-Martin, M.; Sanchez-Esguevillas, A.; Hernandez-Callejo, L.; Arribas, J.I.; Carro, B. Additive ensemble neural network with constrained weighted quantile loss for probabilistic electric-load forecasting. *Sensors* **2021**, *21*, 2979.
40. Shalihin, N.; Firdaus, F.; Yulia, Y.; Wardi, U. Ramadan and strengthening of the social capital of indonesian muslim communities. *HTS Teologiese Studies / Theological Studies* **2020**, *76*. https://doi.org/10.4102/HTS.V76I3.6241.
41. Krawiec, M.; Górska, A. Analysis of Seasonal Patterns in the Performance of Fuel Markets in the Visegrad Group. *Comparative Economic Research* **2024**, *27*. https://doi.org/10.18778/1508-2008.27.12.
42. Marx, D. Success on Repeat. *Package Printing* **2024**, *71*.
43. Hahn, M.; Rofin, M. Why are Sensitive Functions Hard for Transformers? In Proceedings of the Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2024, Vol. 1. https://doi.org/10.18653/v1/2024.acl-long.800.
44. Gagnon, E.; López-Salido, D. Small Price Responses to Large Demand Shocks. *Journal of the European Economic Association* **2020**, *18*. https://doi.org/10.1093/jeea/jvz002.