Leveraging Large Language Models for Discrepancy Value Prediction in Custody Transfer Systems: A Comparative Analysis of Probabilistic and Point Forecasting Approaches

by Fiki Hidayat

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.
Digital Object Identifier 10.1109/ACCESS.2023.1120000

Leveraging Large Language Models for Discrepancy Value Prediction in Custody Transfer Systems: A Comparative Analysis of Probabilistic and Point Forecasting Approaches

FIKI HIDAYAT[1], ARBI HAZA NASUTION[2], FAJRIL AMBIA[1], DIKE FITRIANSYAH PUTRA[1], MULYANDRI[3]
[1]Department of Petroleum Engineering, Universitas Islam Riau, Pekanbaru 28284, Riau, Indonesia
[2]Department of Informatics Engineering, Universitas Islam Riau, Pekanbaru 28284, Riau, Indonesia
[3]PT. Pertamina Hulu Rokan, Pekanbaru 28266, Riau, Indonesia

Corresponding author: Fiki Hidayat (e-mail: fikihidayat@eng.uir.ac.id).

"This work was funded by the Ministry of Education, Culture, Research, and Technology's Direktorat Riset, Teknologi, dan Pengabdian kepada Masyarakat (DRTPM) under grant numbers 112/E5/PG.02.00.PL/2024; 043/LL10/PG..AK/2024; 005/DPPM-UIR/HN-P/2024."

ABSTRACT Discrepancies in custody transfer systems in the oil and gas industry pose significant financial, regulatory, and operational risks. Accurate prediction of these discrepancies is critical to optimizing operations and minimizing potential losses. This study evaluates the effectiveness of Large Language Models (LLMs), specifically the Chronos-FineTuning Amazon Chronos T5 Small model, alongside statistical, machine learning, and deep learning models, in both probabilistic and point forecasting tasks. The evaluation covers metrics such as Weighted Quantile Loss (WQL), Scaled Quantile Loss (SQL), Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and Root Mean Square Error (RMSE). The results highlight the superior performance of the Chronos model in both forecasting paradigms, demonstrating its ability to capture uncertainty and deliver precise predictions. This research offers valuable insights into selecting forecasting methodologies to improve custody transfer operations, underscoring the transformative potential of LLMs in industrial applications.

INDEX TERMS probabilistic time-series forecasting; large language models; discrepancy; custody transfer system

I. INTRODUCTION

Large Language Models, commonly referred to as LLMs, represent sophisticated artificial intelligence systems that have undergone training on vast datasets, enabling them to understand and generate text that closely resembles human-produced language. Built on advanced neural architectures, particularly transformers, LLMs excel at detecting patterns, relationships, and contextual nuances in sequential data [1]. Their flexibility in handling long-term dependencies and ability to adapt to different tasks, such as creative writing and summarizing text, makes them useful for many applications [2]–[5].

In the context of time-series forecasting, the ability of LLMs to process sequential and contextual information makes them particularly effective. Time-series data relies on temporal patterns and dependencies that large language models are meant to detect, just as natural language does. The precise calibration of these models on numerical time-series datasets enables the identification of trends, seasonality, and anomalies, which in turn supports accurate predictions of future values. This capability has wide-ranging applications in fields such as finance, energy, urban planning, and traffic management. For instance, by training on massive datasets containing financial data and media coverage, large language models in finance can identify trends and patterns in mutual fund data, provide insights into mutual fund performance, and create individualized recommendations for mutual funds for individual investors [6]. In the field of energy, large language models have the ability to combine numerical meteorological data with textual weather descriptions, thereby enhancing the degree of precision of forecasts related to photovoltaic power generation. A GPT agent, for example, can interpret linguistic weather descriptions and convert them into numeric vectors, which are then used in conjunction with traditional meteorological data to improve prediction accuracy [7]. Urban planners and traffic managers can benefit from large language models' traffic flow forecasting capabilities, which aid stakeholders in making decisions across a variety of traffic scenarios [8]. The LLMs' capacity to efficiently manage time-series data is a source of various opportunities for informed decision-making and predictive analytics in a diverse array of industries.

In the oil and gas industry, custody transfer is a critical process essential for maintaining accuracy in trade transactions [9]. It involves the precise measurement of hydrocarbons during loading activities. The precise measurement and efficient transfer of hydrocarbon ownership require custody transfer systems (CTS). Discrepancies in these systems, which can arise from a variety of sources, can have a negative impact on financial [10], regulatory [11], and operational outcomes [11]. As a result of these discrepancies, corporations that participate in trade transactions usually suffer significant financial losses.

Discrepancies in custody transfer systems in the oil and gas industry are nearly impossible to eradicate due to the inherent complexities of measurement processes, environmental variability, and equipment limitations. Therefore, in order to reduce possible financial risks, it is crucial to forecast discrepancies in custody transfer systems. The application of advanced technologies and data analytics enables industries to forecast and mitigate discrepancies in a proactive manner, thereby averting potential costs associated with these issues.

A variety of factors can contribute to discrepancies in custody transfer systems. Errors in flow metering [10], sampling [10], and meter proving [12] can lead to inaccurate readings. Challenges in achieving a homogeneous mixture of crude oil and water for accurate sampling [10], [11]. Other factors that can affect flow meters include temperature, pressure, and viscosity [13], [14]. Inadequate operations and maintenance practices can also be a problem [15]. The use of outdated or inappropriate metering equipment and practices can also lead to inaccurate readings [16]. Finally, there are differences in measurement methodologies, such as error-based and uncertainty-based [17], which can affect the results. These multifaceted challenges highlight the importance of predictive measures that not only detect potential discrepancies in real time but also provide actionable insights to resolve them. Companies can increase the accuracy and dependability of custody transfer operations by systematically addressing these categories. Ultimately, this leads to increased efficiency and cost savings.

This study will focus on evaluating the effectiveness of LLMs in predicting production discrepancies by comparing probability-based (uncertainty-based) and point-based (error-based) measurement methodologies. This study intends to offer important insights for improving overall production process efficiency and custody transfer operations by examining the predictive capabilities of LLMs. Through this research, companies will be able to make informed decisions on which measurement methodologies to implement in order to optimize their custody transfer operations. Understanding the predictive capabilities of LLMs allows companies to reduce errors and uncertainties in production discrepancies, resulting in greater accuracy and cost savings. Overall, the findings of this study have the potential to revolutionize how businesses approach custody transfer operations and improve efficiency in their production processes.

II. RELATED WORK

A. LARGE LANGUAGE MODELS IN TIME-SERIES FORECASTING

In the time-series forecasting domain, large language models (LLMs) have demonstrated tremendous promise by utilizing their capabilities in semantic reasoning and sequence modeling to address a variety of challenges associated with time-series forecasting. Gruver et al. [18] explore how LLMs like GPT-3 and Llama-2 can do time-series forecasting using zero-shot and few-shot learning. They have shown that their LLMs can handle various time-series tasks without changing the underlying model by rephrasing forecasting as text generation. Large language models can also integrate numerical time-series data with textual information, thereby improving forecasting accuracy by utilizing additional context [19].

According to Zhang et al. [20], there are five different ways to use LLMs in time series analysis: (1) prompting (input), (2) quantization (tokenization), (3) aligning (embedding), (4) vision as a bridge (LLM stage), and (5) tool integration (output stage). In order to use zero-shot functionality, prompting sees time series as unprocessed text. For quantization to work, numerical data needs to be turned into discrete tokens. This can be done with methods like K-means clustering or vector quantized variational autoencoder (VQ-VAE). Aligning either uses contrastive learning to align time series embeddings with text embeddings or incorporates time series into large language model architectures. Vision as a bridge employs visual representations, such as plots, to synchronize time series with textual data via vision-language models. Tool integration utilizes LLMs indirectly to produce tools (e.g., APIs or code) for designated tasks.

Although LLMs have been progressively utilized in time series forecasting across various fields [21]–[23], including finance, healthcare, and spatio-temporal analysis (e.g., traffic and human mobility) [3], their implementation in the oil and gas industry is still restricted, highlighting considerable unexploited potential. Theoretical frameworks indicate that the adaptability of LLMs renders them particularly effective for intricate predictive tasks within this industry. He Liu et al. [24] investigate the evolution and utilization of sophisticated AI models in the oil and gas sector, encompassing LLMs, Visual Large Models (VLMs), and Multimodal Large Models (MLMs). These models have proven effective in improving efficiency, decision-making, and predictive capabilities in areas such as exploration, drilling, and reservoir management. Nonetheless, present applications in the oil and gas sector are predominantly confined to functions such as intelligent assistance, data analysis, seismic image segmentation, and pipeline fault detection. Prominent instances comprise GeoGPT for geological inquiries, PetroQA for technical document queries, and RockSAM for precise segmentation of geological images. The restricted range of applications underscores the potential to broaden LLM functionalities to time series forecasting, fulfilling essential requirements like production trend prediction, reservoir performance evaluation, and maintenance scheduling, where their capacity to analyze and model sequential data [23] could yield significant advantages.

B. PROBABILISTIC FORECASTING METHODS

Probabilistic time-series forecasting seeks to generate a predictive distribution of the variable of interest rather than a singular point estimate, whereas traditional deterministic forecasting yields a single-value prediction of future outcomes [25]–[27]. Traditional deterministic forecasting methods generally fail to offer the confidence intervals necessary to manage uncertainty, whereas probabilistic forecasting models do [28]–[30]. Traditional deterministic forecasting models rely directly on historical data, whereas probabilistic forecasting models employ Bayesian inference and ensemble methods to produce probabilistic forecasts that incorporate uncertainty in the predictions [31], [32].

Probabilistic forecasts offer a superior depiction of forecast uncertainties, enhance human proficiency in decision-making, and facilitate effective communication of forecast uncertainties to end-users [31], [33]. Probabilistic forecasts assist decision-makers in evaluating risks and making more informed decisions. Probabilistic forecasting effectively quantifies prediction uncertainty in power systems utilizing renewable energy sources [29], [34], offering comprehensive forecasting information and essential data support for analysis and decision-making. Furthermore, cases within the oil and gas industry demonstrate that probabilistic forecasting can improve the reliability of decision-making regarding production predictions for unconventional oil and gas wells [35]. Afifi et al. [35] discovered that employing prediction intervals enhances reliability in comparison to single-value forecasts. The researchers discovered that employing Prediction Interval Coverage Probability (PICP) and Prediction Interval Normalized Average Width (PINAW) establishes a distinct trade-off between interval coverage and precision, facilitating informed decision-making. Maldonado-Cruz and Pyrcz's [36] research delineates the creation and utilization of Temporal Fusion Transformers (TFTs) for forecasting subsurface resources, specifically regarding multi-well fluid flow performance. They can assess uncertainty by employing a quantile loss function to forecast outcome ranges with confidence intervals (e.g., P10 to P90). The TFTs enhance well productivity and resource recovery while considering uncertainties in forecasts.

III. MATERIALS AND METHODS

A. DATASET

The point of this study is to look into how large language models (LLMs) can be used in probabilistic time-series forecasting. More specifically, they will be used to estimate discrepancy values in oil production gathering stations. The dataset utilized in this study was obtained from nine production gathering stations located in the West Area of Riau Province, Indonesia. These stations are Kotabatak, Petapahan, Suram, Kasikan, Terantam, Osam, Langgak, Lindai, and Mahato. For this study, two types of data were collected: the discrepancy gross fraction (DGF) and the discrepancy unallocated net fraction (DUNF). DGF and DUNF differ in their inclusion of a metering factor parameter at DUNF. The discrepancy was determined by comparing the total amount of crude oil delivered by eight gathering stations (Petapahan, Suram, Kasikan, Terantam, Osam, Langgak, Lindai, and Mahato) to the amount of oil recorded at the main station (Kotabatak). Data collection occurred over a span of 650 days, specifically from March 6, 2021, to December 14, 2022, with manual data acquisition conducted daily.

B. METHODS

In this study, we evaluate a diverse set of forecasting models categorized into Statistical Models, Machine Learning Models, Deep Learning Models, and Large Language Models (LLMs). We use the Autogluon [37] Python library for all of our model evaluations. This categorization allows for a comprehensive comparison of traditional and modern approaches to time series forecasting. Each category represents a distinct methodology with strengths tailored to specific data patterns and forecasting tasks.

Forecasting is a challenging task, and no single model consistently outperforms others across all datasets or time series patterns due to variations in trend, seasonality, noise, and dependencies. Using models from multiple categories ensures the evaluation of a wide spectrum of capabilities:

- **Statistical Models** offer simplicity, interpretability, and robustness, especially for stationary or seasonal data.
- **Machine Learning Models** provide flexibility to learn patterns from engineered or automatically extracted features, performing well with non-linear relationships.
- **Deep Learning Models** excel in capturing complex temporal dependencies and large-scale data patterns.
- **Large Language Models (LLMs)** leverage their generalization ability and pre-trained knowledge to forecast without extensive domain-specific tuning.

This multi-model approach provides a nuanced understanding of model performance across diverse forecasting scenarios.

1) Statistical Models

Statistical models are traditional approaches that rely on predefined mathematical structures to represent and forecast time series. These models are particularly effective for capturing simple relationships, seasonality, and trends in the data.

- **AutoETS**: AutoETS [38], [39] automatically selects the best-fitting Exponential Smoothing State Space Model

(ETS) configuration. It is well-suited for handling seasonal and trend components in time series.

- **DynamicOptimizedTheta**: The Theta model [40] decomposes a time series into trend and seasonality components. Its optimized version dynamically adjusts decomposition parameters for better adaptability across different datasets.
- **SeasonalNaive**: SeasonalNaive is a baseline model that assumes future values repeat the observed values from the previous seasonal period. It serves as a reference for evaluating model performance.
- **NPTS**: Non-Parametric Time Series (NPTS) is a kernel-based approach that avoids strong parametric assumptions, making it useful for forecasting time series with irregular patterns.

Each statistical model is tailored for specific data structures, such as seasonality (SeasonalNaive), trend decomposition (DynamicOptimizedTheta), or parameter flexibility (AutoETS, NPTS). Including all ensures comprehensive evaluation. Table 1 shows the parameters for each statistical model.

Model	Parameters
AutoETS	• model = 'ZZZ' • seasonal_period = None • damped = False • n_jobs = 0.5 • max_ts_length = 2500
DynamicOptimizedTheta	• decomposition_type = 'multiplicative' • seasonal_period = None • n_jobs = 0.5 • max_ts_length = 2500
SeasonalNaive	• *seasonal_period = None* • *n_jobs = 0.5*
NPTS	• kernel_type = 'exponential' • exp_kernel_weights = 1.0 • use_seasonal_model = True • num_samples = 100 • num_default_time_features = 1 • n_jobs = 0.5 • max_ts_length = 2500

TABLE 1. Statistical Model Parameters

2) Machine Learning Models

Machine learning models use data-driven methods to learn patterns from time series, enabling them to adapt to non-linear relationships and a variety of features. These models do not rely on predefined structures and instead optimize feature representations to improve forecasting accuracy. Table 2 shows the parameters for each statistical model.

- **TiDE**: TiDE [41] is a Multi-layer Perceptron (MLP)-based encoder-decoder model that uses dense layers to encode input features and predict future values efficiently.
- **RecursiveTabular**: RecursiveTabular iteratively forecasts time steps using tabular regression techniques, building on its previous predictions.
- **DirectTabular**: DirectTabular avoids error propagation by forecasting all future time steps simultaneously using tabular machine learning methods.

By evaluating models like RecursiveTabular and DirectTabular, we assess differences in forecasting strategies (recursive vs. direct). TiDE adds a neural network-based machine learning approach for comparison, highlighting flexibility in capturing non-linearities. Table 2 shows the parameters for each machine learning model.

Model	Parameters
TiDE	• context_length = max(64, 2 * prediction_length) • disable_static_features = False • disable_known_covariates = False • feat_proj_hidden_dim = 4 • encoder_hidden_dim = 64 • decoder_hidden_dim = 64 • dropout = 0.2 • learning rate (lr) = 1e-4 • early stopping patience = 20
RecursiveTabular	• lags = None • date_features = None • target_scaler = 'standard' • max_num_samples = 1,000,000
DirectTabular	• lags = None • date_features = None • target_scaler = 'mean_abs' • max_num_samples = 1,000,000

TABLE 2. Machine Learning Model Parameters

3) Deep Learning Models

Deep learning models capture complex temporal dependencies and patterns in time series data by leveraging hierarchical representations. They are particularly advantageous when dealing with large-scale, high-dimensional datasets. Table 3 shows the parameters for each statistical model.

- **TemporalFusionTransformer**: TemporalFusionTransformer (TFT) [42] integrates attention mechanisms to capture long- and short-term dependencies while effectively combining static and dynamic features.
- **PatchTST**: PatchTST [43] segments input time series data into patches and applies a transformer architecture, allowing the model to capture global and local patterns.
- **DeepAR**: DeepAR [44] is an autoregressive neural network designed for probabilistic forecasting. It outputs probability distributions instead of point predictions, excelling in scenarios with sparse or uncertain data.

Each deep learning model offers unique capabilities: TFT combines static and temporal features, PatchTST emphasizes global temporal patterns, and DeepAR focuses on probabilistic forecasts. Evaluating all three helps identify strengths in different forecasting contexts. Table 3 shows the parameters for each deep learning model.

Model	Parameters
TemporalFusionTransformer	• context_length = max(64, 2 * prediction_length) • hidden_dim = 32 • num_heads = 4 • dropout_rate = 0.1 • max_epochs = 100 • batch_size = 64 • learning rate (lr) = 1e-3
PatchTST	• context_length = 96 • d_model = 32 • nhead = 4 • num_encoder_layers = 2 • dropout_rate = 0.1 • max_epochs = 100 • batch_size = 64 • learning rate (lr) = 1e-3
DeepAR	• context_length = max(10, 2 * prediction_length) • hidden_size = 40 • dropout_rate = 0.1 • max_epochs = 100 • batch_size = 64 • learning rate (lr) = 1e-3

TABLE 3. Deep Learning Model Parameters

4) Large Language Models (LLMs)

Large Language Models (LLMs) apply transfer learning to forecasting tasks, leveraging their pre-trained knowledge to generalize across datasets without extensive fine-tuning. Chronos ZeroShot Bolt Mini [45] uses a zero-shot approach, forecasting directly from input data without prior domain-specific training. This demonstrates the generality and adaptability of LLMs in time series applications. The inclusion of LLMs highlights their potential in forecasting tasks, especially when domain-specific training is limited or when generalization across diverse time series is needed. Table 4 shows the parameters for the Chronos model.

Model	Parameters
Chronos-Bolt-Mini (ZeroShot)	• batch_size = 16 • num_samples = 20 • torch_dtype = 'auto' • data_loader_num_workers = 0 • fine_tune = False

TABLE 4. Large Language Model Parameters

IV. RESULTS AND DISCUSSION

The evaluation of the models was conducted using both probabilistic and point forecasting metrics. Table 5 summarizes the performance of the models across five key metrics: Weighted Quantile Loss (WQL) and Scaled Quantile Loss (SQL) for probabilistic forecasting, and Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and Root Mean Square Error (RMSE) for point forecasting. The models are grouped into four categories: Statistical Models, Machine Learning Models, Deep Learning Models, and Large Language Models (LLMs), providing insights into their comparative performance.

A. PROBABILISTIC FORECASTING

The performance of the models in probabilistic forecasting was evaluated using Weighted Quantile Loss (WQL) and Scaled Quantile Loss (SQL). These metrics provide insights into the models' ability to capture uncertainty in forecasts, which is crucial for risk-sensitive applications.

1) Overview of Results

The results highlight the superior performance of the Chronos ZeroShot Bolt Mini model, particularly in capturing uncertainty and delivering precise predictions. As shown in Figures 1 and 2, this model achieves the lowest Weighted Quantile Loss (WQL) of 0.0322 and Scaled Quantile Loss (SQL) of 0.2604, outperforming both deep learning and machine learning models. This highlights the capability of large language models in probabilistic forecasting tasks, potentially due to their ability to encode complex patterns and relationships in time series data. These results reinforce the model's capability in balancing precision and robustness, making it a highly effective choice for uncertainty-aware forecasting.

Examining the competition performance of other model categories is crucial, even though the Chronos ZeroShot Bolt Mini leads in both SQL and WQL measures. TemporalFusionTransformer outperforms other deep learning models, with a SQL of 0.2706 and a WQL of 0.0335. This is likely attributed to its inherent architecture, which combines LSTM-based encoders with self-attention mechanisms, allowing it to model temporal dependencies effectively. Similarly, PatchTST and DeepAR exhibit competitive performance, with SQL values of 0.3019 and 0.3052, respectively, and WQL values of 0.0374 and 0.0378. These findings show that deep learning models manage to be both adaptable and accurate predictors.

On the other hand, machine learning models such as DirectTabular (SQL: 0.2895, WQL: 0.0358) and TiDE (SQL: 0.3015, WQL: 0.0373) perform similarly to deep learning models. This indicates that structured tabular data approaches have the potential to be viable alternatives to deep learning models, particularly in situations where computational efficiency is a primary concern.

TABLE 5. Model Evaluation

Model	Probabilistic Forecasting		Point Forecasting		
	WQL	SQL	MAE	SMAPE	RMSE
Chronos ZeroShot Bolt Mini	0.032227	0.260369	0.018393	0.038734	0.028986
TemporalFusionTransformer	0.033498	0.270632	0.020489	0.042312	0.028303
DirectTabular	0.035831	0.289482	0.019372	0.040335	0.028388
NPTS	0.037211	0.300631	0.021071	0.043522	0.029296
TiDE	0.037315	0.301476	0.019797	0.041296	0.028844
DeepAR	0.037366	0.301884	0.022098	0.045759	0.030580
PatchTST	0.037778	0.305214	0.020642	0.043407	0.030588
AutoETS	0.042597	0.344142	0.023199	0.047175	0.029623
DynamicOptimizedTheta	0.047442	0.383285	0.022266	0.045447	0.028882
SeasonalNaive	0.090701	0.732779	0.045147	0.090000	0.055393
RecursiveTabular	0.116078	0.937801	0.041380	0.082891	0.051309

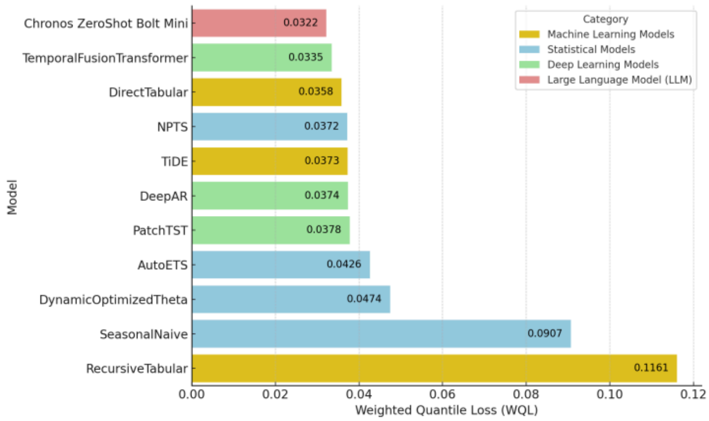

FIGURE 1. Model Performance on WQL.

Statistical models, on the other hand, have observable performance restrictions even though they are interpretable. AutoETS (SQL: 0.3441, WQL: 0.0426) and DynamicOptimizedTheta (SQL: 0.3833, WQL: 0.0474) exhibit inferior performance compared to deep learning and machine learning models. These models rely on predefined assumptions about seasonality and trends, which might limit their adaptability to complex patterns in probabilistic forecasting. SeasonalNaïve (SQL: 0.7328, WQL: 0.0907) and RecursiveTabular (SQL: 0.9378, WQL: 0.1161) exhibit the highest errors, which are indicative of their difficulty in capturing intricate temporal structures. Particularly for RecursiveTabular, it predicts one step at a time and then uses those predictions as input for subsequent steps. Discovering inaccuracies in early projections often leads to an accumulative inaccuracy. Recursive Tabular is only suited for short-term forecasts, which are subsequently iteratively expanded into long-term projections, resulting in less-than-ideal performance.

2) Insights from Probabilistic Forecasting

- LLMs as a Disruptive Approach: The superior performance of the Chronos ZeroShot Bolt Mini underscores the transformative potential of LLMs in time series forecasting. Unlike other models, LLMs are pretrained on diverse data, enabling them to generalize well even in zero-shot scenarios. Their ability to process large amounts of information and learn non-linear relationships contributes to their success in probabilistic forecasting tasks.
- The Role of Deep Learning Architectures: Deep learning models like TemporalFusionTransformer and PatchTST leverage advanced architectures, such as attention mechanisms, to capture long-term dependencies and interactions. These features allow them to offer a robust alternative to LLMs, especially in scenarios where training data is domain-specific and sufficient in quantity.
- Limitations of Traditional Approaches: While statistical models like AutoETS are computationally efficient and interpretable, their reliance on rigid assumptions about

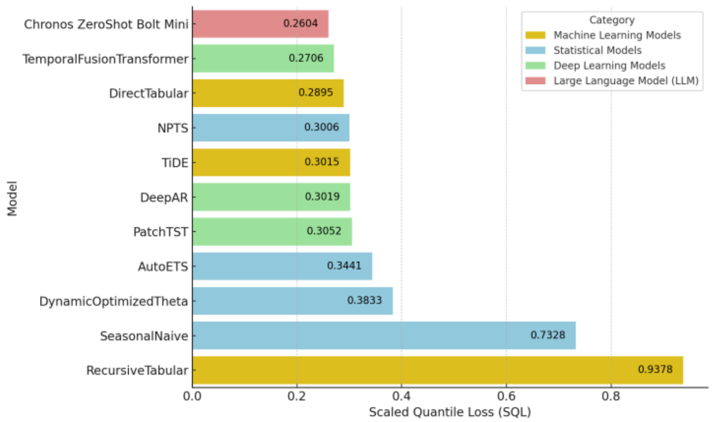

FIGURE 2. **Model Performance on SQL.**

the data limits their effectiveness in complex probabilistic scenarios. Similarly, the SeasonalNaive model's simplistic approach is inadequate for capturing the variability in modern time series data.

- Performance Gaps in Machine Learning Models: Although machine learning models like TiDE and DirectTabular show competitive performance, their results indicate a potential gap in handling uncertainty compared to LLMs and deep learning models. This may suggest that future enhancements in feature engineering and model architecture could narrow the performance disparity.

3) Implications for Applications

Probabilistic forecasting is essential in fields like finance, energy, and supply chain management, where understanding the range of possible outcomes is critical. The results suggest that:

- Chronos ZeroShot Bolt Mini is ideal for applications requiring high precision and the ability to model uncertainty effectively.
- TemporalFusionTransformer provides a balance between performance and computational efficiency, making it suitable for domain-specific forecasting tasks.
- Statistical Models, while suboptimal, may still be valuable in scenarios where interpretability and simplicity are prioritized over accuracy. Possible scenarios include:
  - Small Data Regimes – Statistical models like AutoETS and DynamicOptimizedTheta can be preferable when dataset sizes are too small to effectively train deep learning models [46], [47].
  - Interpretability – Unlike deep learning approaches, statistical models provide clear mathematical formulations, making them useful in regulatory environments where explainability is critical [47].
  - Computational Efficiency — These models are typically faster and require fewer computational resources, making them suitable for real-time forecasting applications with limited processing power [46], [48], [49].

In summary, probabilistic forecasting results demonstrate the significant advantages of LLMs and deep learning models in capturing uncertainty and generating reliable predictions. These findings pave the way for further exploration into how LLMs can be fine-tuned or adapted to enhance their capabilities in domain-specific probabilistic forecasting tasks.

B. POINT FORECASTING

Point forecasting evaluates models based on their ability to provide precise single-point predictions for future values. The metrics used to assess performance—Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and Root Mean Square Error (RMSE)—offer a comprehensive view of prediction accuracy and error magnitude. Here's a detailed discussion of the results for point forecasting.

1) Overview of Results

- Best Performing Model: The Chronos ZeroShot Bolt Mini achieves the lowest MAE of 0.0184, indicating the highest predictive accuracy in absolute terms. Similarly, it attains a low SMAPE of 0.0387, demonstrating its effectiveness in handling percentage-based forecast devia-

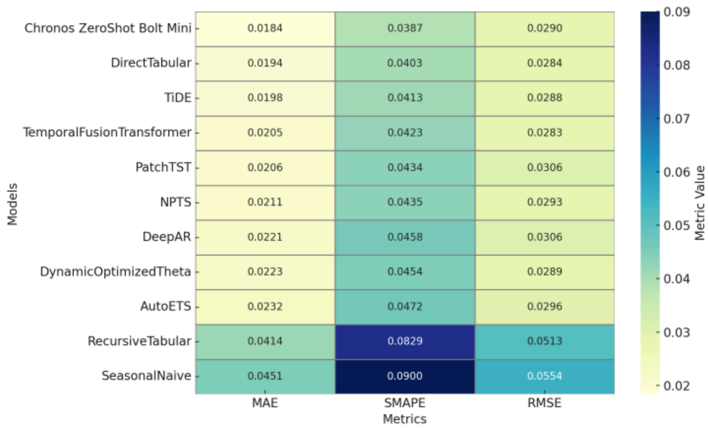

FIGURE 3. Model Performance on Point Forecasting.

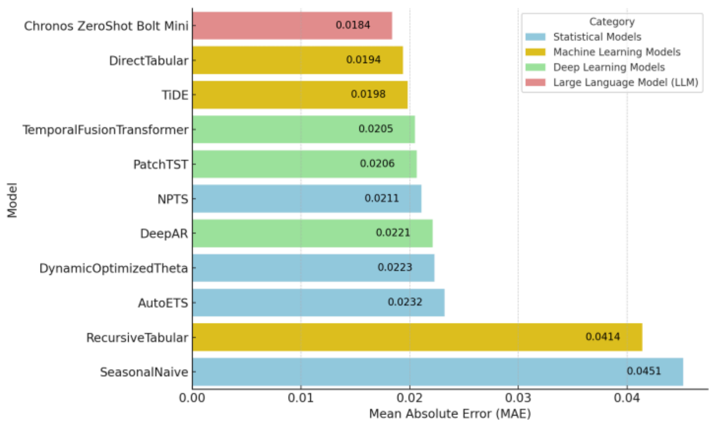

FIGURE 4. Model Performance on MAE.

tions. This underscores the model's superior accuracy in point forecasting, reflecting its ability to capture patterns in time series data with remarkable precision.However, in RMSE, the model performs slightly behind TemporalFusionTransformer (0.0283) and DirectTabular (0.0284) but remains competitive at 0.0290.

- Deep Learning Models: Among deep learning architectures, TemporalFusionTransformer (MAE: 0.0205, RMSE: 0.0283, SMAPE: 0.0423) and PatchTST (MAE: 0.0206, RMSE: 0.0306, SMAPE: 0.0434) emerge as strong contenders. Their performance is comparable to Chronos ZeroShot Bolt Mini, highlighting the adaptability of transformer-based architectures in time-series forecasting. DeepAR also performed competitively, but their slightly higher MAE and SMAPE values highlight areas where they lag behind.

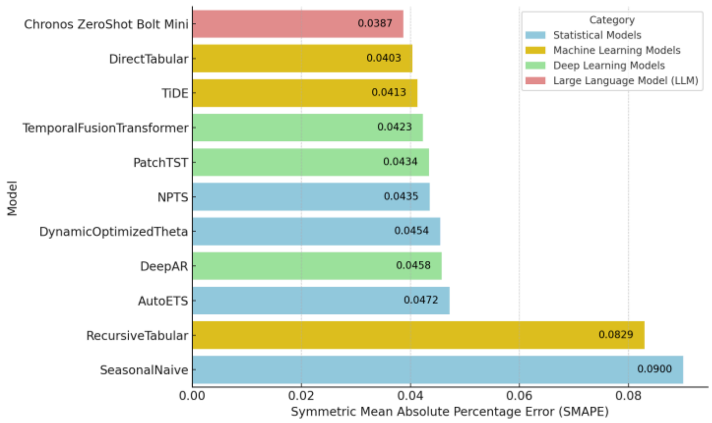

FIGURE 5. Model Performance on SMAPE.

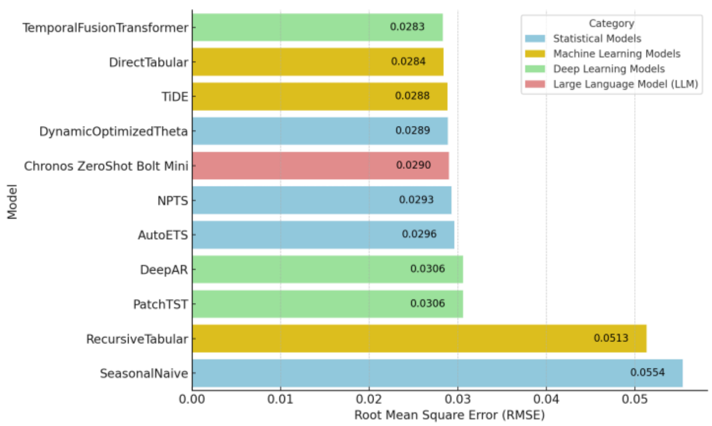

FIGURE 6. Model Performance on RMSE.

- Machine Learning Models: Machine learning models like DirectTabular (MAE: 0.0194, RMSE: 0.0284, SMAPE: 0.0403) and TiDE (MAE: 0.0198, RMSE: 0.0288, SMAPE: 0.0413) balance efficiency and accuracy, making them suitable alternatives to deep learning models in scenarios requiring lower computational costs. These models, however, fell short of the LLM and deep learning models in RMSE, indicating some limitations in handling larger errors.
- Statistical Models: Statistical models, while interpretable, consistently ranked lower in point forecasting metrics. AutoETS and DynamicOptimizedTheta showed moderate performance, but their higher MAE (0.023199 and 0.022266, respectively) and SMAPE values demonstrate limited adaptability to complex patterns. The SeasonalNaive model performed the worst, with substantially higher errors across all metrics (MAE: 0.045147, SMAPE: 0.090000, RMSE: 0.055393), high-

lighting its inadequacy for accurate point forecasting.

2) Key Observations

- LLMs Setting New Standards: The Chronos ZeroShot Bolt Mini's consistent outperformance across all metrics indicates that LLMs are not only adept at capturing uncertainty (probabilistic forecasting) but also excel in point forecasting tasks. This may be due to their ability to leverage extensive pretrained knowledge and adapt to diverse patterns without requiring task-specific training.
- Strengths of Deep Learning Models: Deep learning models like TemporalFusionTransformer and PatchTST demonstrated exceptional performance in metrics such as RMSE and SMAPE. These architectures excel in capturing complex temporal dependencies, which is crucial for accurate point forecasting in dynamic environments.
- Competitive Performance of Machine Learning Models: Models like DirectTabular and TiDE showcased competitive performance, particularly in MAE and SMAPE. Their structured approach to learning relationships in tabular data makes them well-suited for applications where domain knowledge can enhance feature engineering.
- Limitations of Statistical Models: Statistical models such as AutoETS and DynamicOptimizedTheta, despite being reliable in stable environments, struggled with the complexity of real-world time series data. Their higher error rates reflect their inability to adapt to non-linear patterns or unexpected variations.
- SeasonalNaive's Performance: The consistently poor performance of the SeasonalNaive model across all point forecasting metrics underscores its overly simplistic assumptions. By relying solely on repeating past seasonal patterns, it fails to capture the nuances of more complex datasets.

3) Metric-Specific Insights

- Mean Absolute Error (MAE): MAE provides a straightforward measure of prediction accuracy. Chronos's lowest MAE of 0.018393 highlights its precision, while the higher values for SeasonalNaive (0.045147) and RecursiveTabular (0.041380) suggest significant errors in their point forecasts.
- Symmetric Mean Absolute Percentage Error (SMAPE): SMAPE evaluates the relative error in predictions, making it particularly useful for comparing models across datasets with varying scales. The low SMAPE for Chronos (0.038734) and DirectTabular (0.040335) illustrates their robust performance, while SeasonalNaive's SMAPE of 0.090000 reflects its inability to generate accurate predictions.
- Root Mean Square Error (RMSE): RMSE emphasizes larger errors, making it a critical metric for applications sensitive to significant deviations. TemporalFusionTransformer's RMSE of 0.028303 demonstrates its strength in minimizing large prediction errors, while the high RMSE of SeasonalNaive (0.055393) indicates poor accuracy in this regard.

4) Implications for Applications

Point forecasting is essential for industries requiring precise predictions, such as energy demand forecasting, inventory management, and financial market analysis. The results suggest:

- Chronos ZeroShot Bolt Mini is the most suitable choice for applications demanding high accuracy, low error margins, and robust performance across diverse conditions.
- TemporalFusionTransformer is ideal for domains requiring a balance of interpretability, scalability, and strong performance in minimizing larger errors.
- Machine Learning Models like TiDE and DirectTabular can be effective alternatives when computational efficiency and structured data handling are priorities.
- Statistical Models, while interpretable, should be used in environments with limited complexity or when simplicity is preferred over accuracy.

The point forecasting results underscore the dominance of LLMs and deep learning models in delivering accurate predictions. Their advanced architectures and ability to model intricate temporal relationships make them indispensable for complex forecasting tasks. Machine learning models, though competitive, highlight opportunities for improvement in handling larger errors, while statistical models remain limited to simpler use cases. These findings offer a comprehensive view of the trade-offs involved in selecting models for point forecasting.

C. BEST PERFORMING MODEL

The evaluation results revealed that the Chronos ZeroShot Bolt Mini consistently outperformed all other models across both probabilistic and point forecasting metrics, including Weighted Quantile Loss (WQL), Scaled Quantile Loss (SQL), Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and Root Mean Square Error (RMSE). Research by Gruver [18] also supports this since their LLMs show good performance against conventional time-series models. This superior performance highlighted the potential of large language models (LLMs) for accurate time-series forecasting.

Model	Parameters
All Chronos Variants	• batch_size = 32 • fine_tune = True • fine_tune_lr = 2e-4 • fine_tune_steps = 1000 • fine_tune_batch_size = 32

TABLE 6. Chronos FineTuning Parameters

TABLE 7. Fine-tuning Model Evaluation

Model	Size (params)	Probabilistic Forecasting		Point Forecasting		
		WQL	SQL	MAE	SMAPE	RMSE
Chronos-T5-Small	46.2M	-0.029896	-0.241538	-0.018331	-0.037076	0.025458
Chronos-T5-Base	201M	-0.035466	-0.286545	-0.027354	-0.054201	0.038880
Chronos-T5-Tiny	8.39M	-0.036116	-0.291782	-0.022868	-0.046094	0.030291
Chronos-T5-Mini	20.5M	-0.036187	-0.292350	-0.022940	-0.046305	0.030175
Chronos-Bolt-Tiny	8.65M	-0.039267	-0.317245	-0.024236	-0.048840	0.031592
Chronos-Bolt-Mini	21.2M	-0.039449	-0.318710	-0.024534	-0.049370	0.033212
Chronos-Bolt-Base	205M	-0.047734	-0.385644	-0.029538	-0.059342	0.037774
Chronos-Bolt-Small	47.7M	-0.053009	-0.428269	-0.032835	-0.065934	0.040622

Building on this success, we conducted further experiments by fine-tuning all Chronos model variants to explore whether additional performance improvements could be achieved. Fine-tuning allows the model to better adapt to the specific characteristics of the dataset, capturing intricate patterns and reducing prediction errors. Table 6 shows the parameters for the fine-tuning across all Chronos variants. Considering both probabilistic forecasting and point forecasting, Chronos-T5-Small (46.2M parameters) demonstrates the best overall performance among the Chronos and Chronos-Bolt variants as shown in Table 7.

In probabilistic forecasting, Chronos-T5-Small achieves the lowest WQL (-0.029896) and SQL (-0.241538), indicating superior performance in capturing the probabilistic distribution of the time series. As model size increases, WQL and SQL values deteriorate, as seen in Chronos-T5-Base (201M), which has a higher WQL (-0.035466) and SQL (-0.286545), and Chronos-Bolt-Small (47.7M), which has the highest WQL (-0.053009) and SQL (-0.428269). This suggests that smaller models like Chronos-T5-Small can effectively handle probabilistic forecasting without requiring excessive model parameters.

For point forecasting, Chronos-T5-Small achieves the lowest errors across MAE (-0.018331), SMAPE (-0.037076), and RMSE (0.025458), indicating better predictive performance in this category as well. Larger models like Chronos-T5-Base (201M) and Chronos-Bolt-Base (205M) show higher errors, with Chronos-Bolt-Small (47.7M) again exhibiting the highest MAE (-0.032835), SMAPE (-0.065934), and RMSE (0.040622). This indicates that Chronos-Bolt variants, despite their architectural modifications for faster inference, do not provide an advantage in these metrics over the original Chronos-T5 models.

1) Key Observations

- Chronos-T5-Small (46.2M) achieves the best performance across all evaluation metrics, outperforming larger models in both probabilistic (WQL, SQL) and point forecasting (MAE, SMAPE, RMSE).
- Chronos-Bolt variants consistently underperform compared to Chronos-T5 models, with Chronos-Bolt-Small (47.7M) having the worst WQL, SQL, MAE, SMAPE, and RMSE values.
- Increasing model size does not improve performance, as seen in Chronos-T5-Base (201M) and Chronos-Bolt-Base (205M), which show worse results than smaller models like Chronos-T5-Small.
- Among the Chronos-Bolt models, there is no clear advantage in model size scaling, as Chronos-Bolt-Tiny (8.65M) achieves better results than larger Chronos-Bolt variants (Mini, Base, and Small).

Overall, Chronos-T5-Small provides the best trade-off between model size and forecasting performance across all evaluation metrics, making it the most effective model for probabilistic and point forecasting tasks within the Chronos family.

2) Discrepancy Gross Fraction Parameter

The plot in Figure 7 illustrates the observed and forecasted values for a time series using the Chronos-FineTuning Amazon Chronos T5 Small model with the Discrepancy Gross Fraction (DGF) parameter. The following observations and insights are derived from the results:

- **Observed Data (Blue Line)**: The blue line represents the actual observed values of the target variable over time, spanning from August 2022 to December 2022. The observed values demonstrate significant variability, with fluctuations occurring at regular intervals. There are noticeable peaks and troughs, indicating dynamic patterns in the underlying data.
- **Forecasted Data (Orange Line)**: The orange line shows the predictions generated by the Chronos T5 Small model with the DGF parameter. The forecasted line closely follows the observed trend, particularly in the later months of 2022 (October through December). However, there are deviations where the model fails to fully capture the magnitude of some spikes and dips in the observed data.
- **Uncertainty Bands (Shaded Orange Area)**: The shaded region around the forecasted line represents the uncertainty intervals of the model's predictions. These intervals indicate the range within which the model expects the true values to fall with a certain level of confidence. The bands widen slightly in the later months (November and December), suggesting increasing uncertainty as the forecast horizon extends.
- **Model Performance**: The model exhibits a reasonable fit, with the forecasted values aligning well with the overall trend of the observed data. However, there are

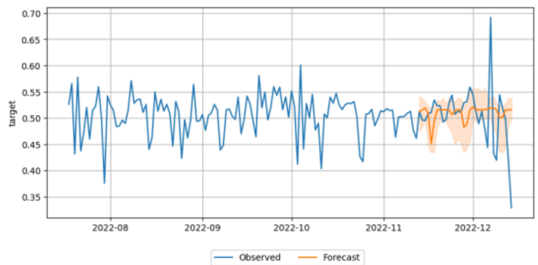

FIGURE 7. Chronos-FineTuning Amazon Chronos T5 Small DGF.

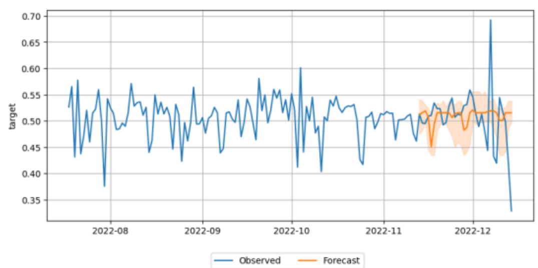

FIGURE 8. Chronos-FineTuning Amazon Chronos T5 Small DUNF.

areas where the model underestimates or overestimates the observed values, particularly during sharp spikes (e.g., early August and late December).

- **Seasonal and Trend Dynamics**: The model appears to capture the general seasonal behavior of the data but struggles with capturing abrupt, non-linear changes in the observed series. This behavior could suggest that while the Chronos T5 Small model with the DGF parameter is adept at modeling trends and smoother variations, its ability to react to rapid, extreme deviations might be limited.

Chronos-FineTuning Amazon Chronos T5 Small model with the DGF parameter demonstrates strong predictive performance for point forecasting and trend estimation. Its close alignment with observed data and the inclusion of uncertainty bands enhance its reliability for real-world applications. However, the model's difficulty in accurately predicting extreme fluctuations suggests that further enhancements, such as incorporating additional contextual features or employing more complex architectures, could improve its accuracy in highly volatile scenarios.

3) Dicrepancy Unallocated Net Fraction Parameter

The plot in Figure 8 shows the observed and forecasted values for a time series using the Chronos-FineTuning Amazon Chronos T5 Small model with the Discrepancy Unallocated Net Fraction (DUNF) parameter. This is compared to the previously discussed results with the DGF parameter (Figure 7). The following observations and insights can be drawn:

- **Observed Data (Blue Line)**: The blue line represents the actual observed values of the target variable from August 2022 to December 2022. Similar to the results with the DGF parameter, the observed data demonstrate significant variability, characterized by frequent fluctuations and occasional extreme spikes, notably the sharp increase at the end of December 2022.
- **Forecasted Data (Orange Line)**: The orange line displays the predictions generated by the Chronos T5 Small model with the DUNF parameter. The forecasted values align well with the general trends in the observed data, particularly during stable periods. However, the model struggles to capture the full magnitude of sharp fluctuations, such as the spike at the end of December. This behavior is consistent with the results for DGF, indicating the inherent difficulty of the model in predicting extreme

deviations.

- **Uncertainty Bands (Shaded Orange Area)**: The shaded region represents the uncertainty intervals for the forecasts. These intervals capture the range within which the model predicts the true values are likely to fall. For the DUNF parameter, the uncertainty bands are slightly wider than those for DGF, reflecting greater uncertainty in the predictions. This is particularly evident in November and December 2022. Despite the widening intervals, some observed values, particularly the extreme spike in December, fall outside the forecast range, indicating that the model underestimates uncertainty for such deviations.
- **Comparative Analysis (DUNF vs. DGF)**:
  - Trend Similarity: Both DUNF and DGF parameters allow the Chronos T5 Small model to effectively capture general trends and seasonal patterns in the observed data.
  - Uncertainty Handling: The DUNF parameter produces slightly wider uncertainty bands than the DGF parameter, indicating higher variance in the forecasts.
  - Extreme Fluctuations: Both parameter configurations struggle to capture the magnitude of sharp spikes, such as the one at the end of December.
- **Insights and Implications**:
  - Strengths: The Chronos T5 Small model demonstrates robustness in capturing overall trends and moderate fluctuations, regardless of the parameter configuration (DGF or DUNF). The use of uncertainty bands provides valuable insights for probabilistic forecasting, helping to assess the confidence level of predictions.
  - Weaknesses: Both DUNF and DGF configurations show limitations in predicting extreme deviations, particularly during periods of high volatility. This is also confirmed by Mulyalim et al. [50], who suggest that data heterogeneity in time-series models frequently complicates the generalizability of LLMs. The model's underestimation of uncertainty during sharp spikes suggests the need for improved mechanisms to handle non-linear and abrupt changes.
  - Applications: The model is well-suited for applications requiring accurate trend forecasting and moderate variability, such as demand prediction, inventory management, and general market analysis. For highly dynamic environments where extreme fluctuations are common, additional refinements or hybrid approaches may be necessary.

The results for the DUNF parameter show similar performance trends to those for the DGF parameter. While the Chronos-FineTuning Amazon Chronos T5 Small model effectively captures overall trends and seasonal patterns, its difficulty in handling extreme deviations remains evident. Wider uncertainty bands in the DUNF configuration highlight increased variability, providing an additional layer of interpretability for the forecasts. Future enhancements could focus on improving the model's response to non-linear behaviors and extreme values to extend its applicability to more dynamic settings.

V. CONCLUSIONS

This study comprehensively evaluated the performance of various forecasting models, including Large Language Models (LLMs), deep learning models, machine learning models, and statistical models, for predicting discrepancies in custody transfer systems. The Chronos-FineTuning Amazon Chronos T5 Small model consistently outperformed other models across probabilistic and point forecasting metrics, showcasing its robustness in capturing uncertainty and delivering precise predictions.

Deep learning models, such as the TemporalFusionTransformer, demonstrated competitive performance, particularly in minimizing errors during point forecasting tasks. Machine learning models like TiDE and DirectTabular provided an effective balance of flexibility and accuracy but exhibited limitations in handling uncertainty. Statistical models, while interpretable and efficient, struggled to adapt to complex patterns and exhibited higher error rates compared to modern approaches.

The findings emphasize the transformative potential of LLMs in industrial time-series forecasting, particularly in mitigating risks associated with custody transfer discrepancies. Future work could focus on enhancing the model's ability to handle extreme deviations and incorporating domain-specific features to further improve predictive accuracy. This research underscores the importance of leveraging advanced AI-driven methodologies to revolutionize operational efficiency in the oil and gas industry.

REFERENCES

[1] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv*, 7 2023.

[2] Seyed Kourosh Mahjour, Ramin Soltanmohammadi, Ehsan Heidaryan, and Salah A. Faroughi. Geosystems risk and uncertainty: The application of chatgpt with targeted prompting. *Geoenergy Science and Engineering*, 238, 7 2024.

[3] Gang Jiang, Zhihao Ma, Liang Zhang, and Jianli Chen. Eplus-llm: A large language model-based computing platform for automated building energy modeling. *Applied Energy*, 367, 8 2024.

[4] Xingzhi Wang, Nabil Anwer, Yun Dai, and Ang Liu. Chatgpt for design, manufacturing, and education. In *Procedia CIRP*, volume 119, pages 7–14. Elsevier B.V., 2023.

[5] Qian Wei, Hongjun Sun, Yin Xu, Zisheng Pang, and Feixiang Gao. Exploring the application of large language models based ai agents in leakage detection of natural gas valve chambers. *Energies*, 17, 11 2024.

[6] Mrunali Ghodake, Pranjal Misal, Tanvi Tendulkar, and Swapnaja Hiray. Mutual fund analysis using large language model. In *2024 MIT Art, Design and Technology School of Computing International Conference, MITADTSoCiCon 2024*, 2024.

[7] Ziming Yan, Du Zhenyuan, Yan Xu, and Ziyan Zhou. Probabilistic pv power forecasting by a multi-modal method using gpt-agent to interpret weather conditions. In *2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)*, pages 1–6. IEEE, 2024.

[8] Xusen Guo, Qiming Zhang, Junyue Jiang, Mingxing Peng, Meixin Zhu, and Hao Frank Yang. Towards explainable traffic flow prediction with large language models. *Communications in Transportation Research*, 4:100150, 12 2024.

[9] Nurul Shaidatul Shima Mohamad Daud, Datu Darul Anwar Datu Shahrun, and Nur Melatee Mohd Fauzi. Elma (export loading monitoring & analysis): Transformative innovations in custody transfer monitoring and analysis at petronas export terminals. In *Abu Dhabi International Petroleum Exhibition and Conference*, page D041S136R003. SPE, 2024.

[10] W. Maru, S. Lakshmanan, N. Singh, and A. Thomas. Oil-water flow measurement for custody transfer applications. In *FLOMEKO 2019 - 18th International Flow Measurement Conference*, 2019.

[11] Gary Potten. Realizing value from adherence to crude oil custody transfer automatic sampling standards. In *Society of Petroleum Engineers - SPE Nigeria Annual International Conference and Exhibition 2021, NAIC 2021*, 2021.

[12] Jasmina Perisic, Marina Milovanovic, Ivana Petrovic, Ljiljana Radovanovic, Marko Ristic, James G. Speight, and Velimir Perisic. Application of a master meter systemto assure crude oil and natural gas quality during transportation. *Petroleum Science and Technology*, 36(16):1222 – 1228, 2018.

[13] Hao Zhu, Alfred Rieder, Paul Ceglia, and Hai-Thuy Ngo. Laboratory and field validation of a new coriolis metering concept for better measurement uncertainty, reliability and process insight. *Measurement: Journal of the International Measurement Confederation*, 218, 2023.

[14] A. Rieder, H. Zhu, P. Ceglia, and H.-T. Ngo. Laboratory and field validation of a new coriolis metering concept for better measurement uncertainty, reliability and process insight. In *19th International Flow Measurement Conference 2022, FLOMEKO 2022*, 2022.

[15] M. Aslam, K. Venkatesan, S. Ahmed, A. Al Ketbi, M.S. Ahmed, K. Alshehhi, F. Haider, and M.K. Al Mazrouei. Session: 47 | health and safety: Human factors and process safety. In *Society of Petroleum Engineers - ADIPEC, ADIP 2023*, 2023.

[16] Michael Iwegbu, Dulu Appah, and Emma Agu Ogwo. Crude oil metering experience in the niger-delta. In *Society of Petroleum Engineers - Nigeria Annual International Conference and Exhibition 2007, NAICE 2007*, 2007.

[17] E.A. Shalovnikov, A.N. Krasnov, and M. Yu Prakhova. Estimation of the accuracy of measurements of oil mass in fiscal metering. *Journal of Physics: Conference Series*, 1889(5), 2021.

[18] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *NIPS'23: 37th International Conference on Neural Information Processing Systems*, pages 19622–19635, 12 2023.

[19] Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23343–23351, 3 2024.

[20] Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K. Gupta, and Jingbo Shang. Large language models for time series: A survey. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, pages 8335–8343. International Joint Conferences on Artificial Intelligence Organization, 8 2024.

[21] Wenlong Liao, Shouxiang Wang, Dechang Yang, Zhe Yang, Jiannong Fang, Christian Rehtanz, and Fernando Porté-Agel. Timegpt in load forecasting: A large time series model perspective. *Applied Energy*, 379, 2 2025.

[22] Hua Tang, Chong Zhang, Mingyu Jin, Qinkai Yu, Zhenting Wang, Xiaobo Jin, Yongfeng Zhang, and Mengnan Du. Time series forecasting with llms: Understanding and enhancing model capabilities. *arXiv*, 2 2024.

[23] Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. Position: What can large language models tell us about time series analysis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22260–22276. PMLR, 21–27 Jul 2024.

[24] He LIU, Yili REN, Xin LI, Yue DENG, Yongtao WANG, Qianwen CAO, Jinyang DU, Zhiwei LIN, and Wenjie WANG. Research status and application of artificial intelligence large models in the oil and gas industry. *Petroleum Exploration and Development*, 51:1049–1065, 8 2024.

[25] Marie Courbariaux, Pierre Barbillon, and Éric Parent. Water flow probabilistic predictions based on a rainfall–runoff simulator: a two-regime model with variable selection. *Journal of Agricultural, Biological and Environmental Statistics*, 22:194–219, 6 2017.

[26] Yunyi Zhou, Zhixuan Chu, Yijia Ruan, Ge Jin, Yuchen Huang, and Sheng Li. ptse: A multi-model ensemble method for probabilistic time series forecasting. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4684–4692. International Joint Conferences on Artificial Intelligence Organization, 8 2023.

[27] E. Camporeale, X. Chu, O. V. Agapitov, and J. Bortnik. On the generation of probabilistic forecasts from deterministic models. *Space Weather*, 17:455–475, 3 2019.

[28] Van Kwan Zhi Koh, Xinyi Li, Zhiping Lin, Ye Li, Ehsan Shafiee, and Bihan Wen. Deterministic to probabilistic time series forecasting by recurrent neural network with variational residual boosting. In *2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA)*, pages 203–208, 2023.

[29] W A N Can, C U I Wenkang, and SONG Yonghua. Probabilistic forecasting for power systems with renewable energy sources: Basic concepts and mathematical principles. *Proceedings of the CSEE*, 41:6493–6508, 2021.

[30] I.K. Bazionis, P.A. Karafotis, and P.S. Georgilakis. A comprehensive review in wind power probabilistic forecasting. In *The 12th Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MEDPOWER 2020)*, volume 2020, pages 455–460, 2020.

[31] Ezio Todini. Flood forecasting and decision making in the new millennium. where are we? *Water Resources Management*, 31:3111–3129, 8 2017.

[32] W. Tao, A. Atencia, Y. Li, X. Qiu, Z. Kang, and Y. Wang. Comparison of statistical ensemble methods for probabilistic nowcasting of precipitation. *Meteorologische Zeitschrift*, 29:203–217, 2020.

[33] Rémy Garçon, Benoît Houdant, Federico Garavaglia, Thibault Mathevet, Emmanuel Paquet, and Joël Gailhard. Human assessment of hydrometeorological forecasts and communication of their uncertainties in a decision making context; [expertise humaine des prévisions hydrométéorologiques et communication de leurs incertitudes dans un contexte décisionnel]. *Houille Blanche*, page 71 – 80, 2009.

[34] O Lindberg, R Zhu, and J Widén. Quantifying the value of probabilistic forecasts when trading renewable hybrid power parks in day-ahead markets: A nordic case study. *Renewable Energy*, 237:121617, 2024.

[35] Hadeel Afifi, Mohamed Elmahdy, Motaz El Saban, and Mervat Abu-Elkheir. Probabilistic time series forecasting for unconventional oil and gas producing wells. In *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 450–455, 2020.

[36] Eduardo Maldonado-Cruz and Michael J. Pyrcz. Multi-horizon well performance forecasting with temporal fusion transformers. *Results in Engineering*, 21, 3 2024.

[37] Oleksandr Shchur, Caner Turkmen, Nick Erickson, Huibin Shen, Alexander Shirkov, Tony Hu, and Yuyang Wang. AutoGluon-TimeSeries: AutoML for probabilistic time series forecasting. In *International Conference on Automated Machine Learning*, 2023.

[38] Rob J Hyndman, Anne B Koehler, Ralph D Snyder, and Simone Grose. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18:439–454, 2002.

[39] Rob Hyndman, Anne Koehler, Keith Ord, and Ralph Snyder. *Forecasting with Exponential Smoothing*. Springer Berlin Heidelberg, 2008.

[40] V. Assimakopoulos and K. Nikolopoulos. The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, 16:521–530, 10 2000.

[41] Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023.

[42] Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: A survey, 4 2021.

[43] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

[44] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36:1181–1191, 7 2020.

[45] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Syndar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

[46] Oleksandr Shchur, Ali Caner Turkmen, Nick Erickson, Huibin Shen, Alexander Shirkov, Tony Hu, and Bernie Wang. Autogluon–timeseries: Automl for probabilistic time series forecasting. In Aleksandra Faust, Roman Garnett, Colin White, Frank Hutter, and Jacob R. Gardner, editors, *Proceedings of the Second International Conference on Automated*

Machine Learning, volume 224 of *Proceedings of Machine Learning Research*, pages 9/1–21. PMLR, 12–15 Nov 2023.

[47] Syeda Sitara Wishal Fatima and Afshin Rahimi. A review of time-series forecasting algorithms for industrial manufacturing systems. *Machines*, 12(6), 2024. Cited by: 3; All Open Access, Gold Open Access.

[48] Polina Pilyugina, Svetlana Medvedeva, Kirill Mosievich, Ilya Trofimov, Alina Kostromina, Dmitry Simakov, and Evgeny Burnaev. A large-scale empirical study of aligned time series forecasting. *IEEE Access*, 12:131100 – 131121, 2024. Cited by: 0; All Open Access, Gold Open Access.

[49] Kashif Rasul, Andrew Bennett, Pablo Vicente, Anderson Schneider, Yuriy Nevmyvaka, Umang Gupta, and Hena Ghonia. Vq-tr: Vector quantized attention for time series forecasting. In *12th International Conference on Learning Representations, ICLR 2024*, 2024. Cited by: 0.

[50] Ozan Baris Mulayim, Pengrui Quan, Liying Han, Xiaomin Ouyang, Dezhi Hong, Mario Bergés, and Mani Srivastava. Are time series foundation models ready to revolutionize predictive building analytics? In *BuildSys 2024 - Proceedings of the 2024 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, page 169 – 173, 2024. Cited by: 0.

FIKI HIDAYAT is a Senior Lecturer of the Department of Petroleum Engineering, Universitas Islam Riau, Indonesia. He got his master's degree in Energy and Mineral Resources Engineering from Sejong University, South Korea, in 2015 and obtained a Bachelor's Degree in Petroleum Engineering from Institut Teknologi Bandung, Indonesia, in 2011. His current research is focused on the utilization of machine learning in the oil and gas industry. His research interests also include CO2 storage and flow assurance. His current research on the utilization of machine learning in the oil and gas industry demonstrates his commitment to advancing technology in this field. Additionally, his interest in CO_2 storage and flow assurance highlights his dedication to addressing important environmental and operational challenges within the industry.

ARBI HAZA NASUTION is an Associate Professor of the Department of Informatics Engineering, Universitas Islam Riau, Indonesia. He got his Ph.D. Degree in Informatics from Kyoto University in 2018, obtained Bachelor Degree in Computer Science from National University of Malaysia in 2010 and obtained Master Degree in Management Information System from National University of Malaysia in 2012. His current research interests include computational linguistics, natural language processing, machine learning, and knowledge representation. Currently, he is working on Indonesia Language Sphere Project which aims to semi-automatically create bilingual dictionaries among various Indonesian Ethnic Languages to preserve these languages.

FAJRIL AMBIA is a Reservoir Simulation Specialist and adjunct professor at Department of Petroleum Engineering, Universitas Islam Riau, Indonesia. He earned his Doctorate degree in Petroleum Engineering from Institut Teknologi Bandung in 2017, following the completion of his Master's degree in the same field in 2011 and a Bachelor's degree in Physics, specializing in Earth Physics, in 2009. Fajril's research is centered on reservoir simulation and optimization, with a particular emphasis on incorporating machine learning techniques into subsurface analysis. He plays an active role in the development and implementation of industry standards for oil and gas resources and reserves reporting in Indonesia, demonstrating his commitment to advancing the sector through both academic and practical contributions.

DIKE FITRIANSYAH PUTRA is a senior lecturer of the Department of Petroleum Engineering, Universitas Islam Riau. He is professional engineer with over 24 years' experience internationally in the oil and gas industry (Caltex, Chevron, Statoil, QP, ADCO, BrentEnergy Canada, OGPC UK, Petronas), extensively exposure to complex crude oil and gas in clastic and carbonate. He has proven abilities in dynamic simulation model and surveillance management of reservoir. He has an excellent in Strategic Management implementation and business innovation. He certified by APEGA Canada as Professional Engineer (P.Eng) designation and certified by PII Indonesia as Professional Engineer (Ir.) designation. Currently, he holds position as the chairman of the Centre of Energy Studies and the laboratory of integrated research and testing of Universitas Islam Riau. He was graduated from Bachelor of Science in Chemical Engineering, UGM Yogyakarta. He got duo master in petroleum and business engineering in NTNU Trondheim Norway, and Manchester University UK. Recently, he is a candidate doctor for specific reservoir subject (electrokinetic EOR monitoring) in UTM Johor Malaysia.

MULYANDRI is a Senior Information Technology and Oil Production Operation practitioner. He has experties on big data evaluation, Operational Data Store, Datawarehouse Development and Analysis, and Quantity Assurance Implementation. He got his Master's Degree in Management of Information Technology from Institut Sepuluh Nopember, Indonesia, in 2006, and obtain a Bachelor's Degree in Informatics Engineering from Institut Teknologi Bandung, Indonesia, in 1996. He has been working in several big oil companies in Indonesia for 27 years.

• • •

Leveraging Large Language Models for Discrepancy Value Prediction in Custody Transfer Systems: A Comparative Analysis of Probabilistic and Point Forecasting Approaches

ORIGINALITY REPORT

9%	5%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Recent Trends in Image Processing and Pattern Recognition", Springer Science and Business Media LLC, 2024 Publication	1%
2	arxiv.org Internet Source	1%
3	www.researchgate.net Internet Source	1%
4	Submitted to Adama Science and Technology University Student Paper	1%
5	www.mdpi.com Internet Source	<1%
6	iris.polito.it Internet Source	<1%
7	vbn.aau.dk Internet Source	<1%
8	Adil Ahmed, Muhammad Khalid. "A Nonlinear Autoregressive Neural Network Model for Short-Term Wind Forecasting", 2017 9th IEEE-GCC Conference and Exhibition (GCCCE), 2017 Publication	<1%
9	www.pcsee.org Internet Source	<1%
10	Joshua Yindenaba Abor, Amin Karimu. "Sustainability Management in the Oil and	<1%

Gas Industry - Emerging and Developing Country Perspectives", Routledge, 2023
Publication

11	Submitted to Bournemouth University Student Paper	<1 %
12	Olakunle Elijah, Pang Ai Ling, Sharul Kamal Abdul Rahim, Tan Kim Geok et al. "A Survey on Industry 4.0 for the Oil and Gas Industry: Upstream Sector", IEEE Access, 2021 Publication	<1 %
13	He LIU, Yili REN, Xin LI, Yue DENG, Yongtao WANG, Qianwen CAO, Jinyang DU, Zhiwei LIN, Wenjie WANG. "Research status and application of artificial intelligence large models in the oil and gas industry", Petroleum Exploration and Development, 2024 Publication	<1 %
14	Submitted to Liverpool John Moores University Student Paper	<1 %
15	Emre Bulut. "chapter 10 Market Volatility and Models for Forecasting Volatility", IGI Global, 2024 Publication	<1 %
16	Song Ding, Zhijian Cai, Yanzu Wu, Huahan Zhang, Xingao Shen. "FNNGM: A neural-driven fractional-derivative multivariate fusion model for interpretable real-time CPI forecasts", Knowledge-Based Systems, 2024 Publication	<1 %
17	ebin.pub Internet Source	<1 %
18	Pronaya Bhattacharya, Vivek Kumar Prasad, Ashwin Verma, Deepak Gupta et al.	<1 %

"Demystifying ChatGPT: An In-depth Survey of OpenAI’s Robust Large Language Models", Archives of Computational Methods in Engineering, 2024
Publication

19	eartharxiv.org Internet Source	<1 %
20	engj.org Internet Source	<1 %
21	export.arxiv.org Internet Source	<1 %
22	medium.com Internet Source	<1 %
23	Fitri Amelia, Hesty Parbuntari, Iryani, Ikhwan Resmala Sudji, Sherly Rahmayani, Andini Novita Ramadhani, Shilvira Ananda. "Phytoconstituents of Terminalia catappa linn fruits extract exhibit promising antidiabetic activities against α-amylase and α-glucosidase in vitro and in silico", Informatics in Medicine Unlocked, 2024 Publication	<1 %
24	docslib.org Internet Source	<1 %
25	ojs.aaai.org Internet Source	<1 %
26	"Advances in Information Retrieval", Springer Science and Business Media LLC, 2024 Publication	<1 %
27	F. Fathurrahman, Siti Zahrah, E. Ernita, H. Heriyanto, Imam Mahadi. "The Effects of the Growth Regulator Paclobutrazol on Physiological Characteristics of Rain Tree (*Albizia saman* Jacq. Merr.)", Journal of Ecological Engineering, 2023	<1 %

Publication

28 Miao He, Weiwei Jiang, Weixi Gu. "TriChronoNet: Advancing electricity price prediction with Multi-module fusion", Applied Energy, 2024 <1%
Publication

29 Poler, R.. "Forecasting model selection through out-of-sample rolling horizon weighted errors", Expert Systems With Applications, 201111/12 <1%
Publication

30 T. Mariprasath, Kumar Reddy Cheepati, Marco Rivera. "Practical Guide to Machine Learning, NLP, and Generative AI: Libraries, Algorithms, and Applications", River Publishers, 2024 <1%
Publication

31 Yessenia Y. Montoya, Carlos G. Pillajo, Jessica S. Ortiz. "Training Assistant for LACT Process Through Augmented Reality", 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020 <1%
Publication

32 mdpi.com <1%
Internet Source

33 www.lokad.com <1%
Internet Source

34 www.scirp.org <1%
Internet Source

35 Feiyu Chen, Linghui Sun, Boyu Jiang, Xu Huo, Xiuxiu Pan, Chun Feng, Zhirong Zhang. "A Review of AI Applications in Unconventional Oil and Gas Exploration and Development", Energies, 2025 <1%
Publication

36	Hadeel Afifi, Mohamed Elmahdy, Motaz El Saban, Mervat Abu-Elkheir. "Probabilistic Time Series Forecasting for Unconventional Oil and Gas Producing Wells", 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), 2020 Publication	<1%
37	Sai Kiran Oruganti, Dimitrios A Karras, Srinesh Singh Thakur, Janapati Krishna Chaithanya, Sukanya Metta, Amit Lathigara. "Digital Transformation and Sustainability of Business", CRC Press, 2025 Publication	<1%
38	Song, Zhengming. "Spatial-Temporal Multivariate Time Series Forecasting Using Graph Neural Networks, With an Application to Traffic Speed Prediction.", The Claremont Graduate University Publication	<1%
39	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	<1%
40	Zhang, Xiyuan. "Unifying Physics and Semantics for Robust Sensor Time Series Analysis", University of California, San Diego, 2024 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		