INSPIRE INTELLIGENCE

Inspire Intelligence Journal
Vol. 1, No. 1, 2026
https://doi.org/10.65718/inspireAI.2026.1003

InspireQuill.org
Inspire Academic Press Inc. (Canada)

Research Paper

Prophet and GPT-2 Algorithms for Demand Forecasting in Last-Mile Delivery: A Comparative Analysis and Optimization

Hiba Mabkhout [1, *] and Jamal Benhra [2]

[1] School of Textile and Clothing industries (ESITH), Casablanca; Morocco
[2] LARILE, ENSEM, University Hassan II of Casablanca, Morocco
[*] Corresponding Author: jbenhra@ensem.ac.ma

Abstract

Demand forecasting is a central issue in last-mile logistics, where operational performance depends on a delicate balance between prediction accuracy and interpretability of results. This study offers a rigorous comparison between two radically different modeling approaches: Prophet, a transparent additive model well-suited to decomposable time series, and GPT-2, a modified transformer capable of capturing complex and nonlinear demand dynamics. Using real-world delivery data, we conduct a thorough evaluation of both models based on several criteria: forecast accuracy (MAE/RMSE), robustness to demand volatility, computational efficiency, and sustainability impact. To ensure a fair comparison, hyperparameter optimization is systematically conducted using Optuna, ensuring optimal configurations for each model. The results reveal that Prophet excels in stable demand environments, thanks to its ease of interpretation and regularity, while GPT-2 demonstrates a marked adaptability to unpredictable variations, at the cost of higher computational costs. The article concludes with practical recommendations for logistics stakeholders, proposing tailored selection criteria (accuracy, explainability, resource constraints) based on the specificities of operational environments. This work aims to bring theoretical advances in predictive modeling closer to the realities of last-mile logistics, providing a concrete decision-making framework for implementing forecasting solutions based on artificial intelligence.

Keywords: Prophet; GPT-2; Time series; Last mile logistics; Demand forecasting; Artificial intelligence; Logistics optimization

1. Introduction

Last mile logistics, a critical step representing 15 to 40% of total logistics costs and contributing significantly to CO₂ [1], requires accurate forecasting models to optimize fleets and routes. Traditional methods (ARIMA, linear regression) struggle to integrate complex variables such as weather or promotional events, limiting their usefulness in dynamic contexts [8]. Faced with these challenges, two approaches are emerging: Prophet, a hybrid model designed for time series with explicit [2], and GPT-2, a language model adapted to time sequences via its *transformer architecture* [9]. This paper compares these two models on delivery data of different sizes, evaluating their accuracy (MAE, RMSE) and their impact on fleet sizing and carbon emissions. The objective is to determine

Academic Editor:
Ghazanfar Latif

Received: 28/05/2025
Revised: 16/08/2025
Accepted: 14/11/2025
Published: 29/12/2025

Citation
Mabkhout, H., & Benhra, J. (2026). Prophet and GPT-2 algorithms for demand forecasting in last-mile delivery: A comparative analysis and optimization. *Inspire Intelligence Journal, 1*(1), 29-41.

whether LLMs, despite their complexity, outperform specialized tools like Prophet in a constrained operational context, where interpretability and speed are crucial [5]. The results will guide practitioners in choosing a model that balances performance, costs and sustainability.

This study explores two approaches to customer demand forecasting, implementing the GPT-2 and Prophet models. Each model is optimized through hyperparameter tuning with Optuna and then evaluated using performance metrics for accuracy and efficiency. The performance of different models is compared based on metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The analysis highlights the ability of each model to handle different demand patterns. The study concludes with recommendations for future research and practical applications. In particular, it highlights the potential of Transformer-based models to improve forecasting accuracy and optimize supply chain efficiency. The goal is to provide companies and researchers with a guide for choosing the most appropriate fore-casting methods and strengthening the responsiveness of supply chains to market fluctuations.

2. Background Literature

Prediction models in last mile logistics fall into two categories, each addressing the specific challenges of the sector differently. ARIMA [10] and LSTM [11] are two classical forecasting methods that have been used in many fields and have been the most popular for a long time. Nonetheless, when utilized in last-mile logistics, these models demonstrate numerous significant constraints. ARIMA is good for stationary time series, but it can't include outside factors like weather, marketing campaigns, or special events, which makes it less useful. This limitation makes it much less useful in logistics settings that are very dynamic and where demand is affected by many outside factors. On the other hand, LSTM models can learn complicated non-linear relationships in time series data. However, they need a lot of historical data to work well and are hard to understand. This lack of transparency makes it hard for logistics decision-makers to explain and check their forecasting choices, especially in industries with strict rules [12]. As a result of these problems, newer forecasting models have been suggested as possible solutions. Prophet is a tool made specifically for predicting time series. It uses an explicit additive decomposition framework to separate trend, seasonality, and regressor effects. This means that it can easily include things like holidays and promotions. Prophet's clear structure makes it a great fit for operational logistics settings. It has also shown great performance in urban warehouse settings, with an average forecasting error of about 8% on multivariate datasets [2]. At the same time, fine-tuning has made it possible for large language models like GPT-2 to be used for time series forecasting by treating time series as tokenized text. GPT-2 can model long-range dependencies and capture complex temporal patterns thanks to its transformer-based architecture. However, this comes at the cost of making the model harder to understand and more computationally intensive [7].

The choice of the right forecasting model is also affected by important operational problems that come with last-mile logistics. One big problem is that demand can suddenly spike for reasons like flash sales or bad weather. Because they depend on historical linear patterns [2], classical models like ARIMA always underestimate these sudden changes. On the other hand, more complicated models like LSTM and GPT-2 may overfit when there isn't enough training data. Another important thing to think about is how easy it is to understand, which is especially important in fields like food and drug logistics, where following the rules and being able to trace decisions are important. In this case, Prophet has a clear advantage because it breaks down forecast components in a way that is easy to understand and trust, which helps stakeholders understand and trust the model's outputs. On the other hand, big language models like GPT-2 mostly work as black boxes, which makes it hard to explain or justify their predictions in real-world decision-making situations [5].

Models like Prophet or SARIMA remain popular for their simplicity and interpretability. They are optimal for time series with explicit seasonal patterns and limited data volumes [8]. Hybrid approaches like ARIMA-LSTM combine the statistical rigor of classical models with the flexibility of neural networks, improving accuracy on complex data [13].

Transformers, initially designed for NLP, are adapted to time series via architectures such as Informer [14-15] or TFT [16]. These models capture long-term dependencies and multivariate interactions, ideal for dynamic data. LLMs such as GPT-2, GPT-3 are emerging as versatile tools, capable of merging textual data (e.g. customer reviews) with time series for contextual forecasting [17]**.** A 10% improvement in forecast accuracy can reduce logistics costs by 5–15% by avoiding fleet overutilization and delay penalties [18]. Hybrid models (e.g. Prophet + Gradient Boosting) have enabled companies like Amazon to reduce their storage costs by 12% while maintaining a service rate of 98% [19].

Route optimization through accurate forecasting reduces empty kilometers traveled, reducing CO_2 emissions by 8 to 20% depending on urban density [20]. A pilot application of generative AI (GPT-3) in logistics planning reduced emissions by 15% for a delivery network in Sweden, by synchronizing routes with weather forecasts [21]. The energy cost of deep learning models (e.g., GPT-3 training consumes 1.287 MWh) offsets their ecological gains [22]. Large-scale adoption requires cloud infrastructure and technical skills, a barrier for SMEs [23].

3. Methodology

3.1. Data sets and data preprocessing

In this study, three distinct datasets were used to evaluate the robustness of the two models (GPT-2 and Prophet) to different time series lengths: A 2-year set (~730 rows), A 4-year set (~1460 rows), A 6-year set (~2190 rows) as shown in Figure 1-3. Preprocessing consisted of several crucial steps to ensure the quality and relevance of the models' input data: Data cleaning, Temporal alignment (time series were synchronized according to the same temporal granularity, daily aggregation was applied), additivity and stationarity study: visual tests (plotting of series, display of the rolling mean and rolling standard deviation) and statistics (Dickey-Fuller test) were used to assess the stationarity of the series. Time series decomposition: each series was decomposed into trend, seasonality and residuals using classic tools. ACF (Autocorrelation Function) functions were used to detect significant temporal dependencies.

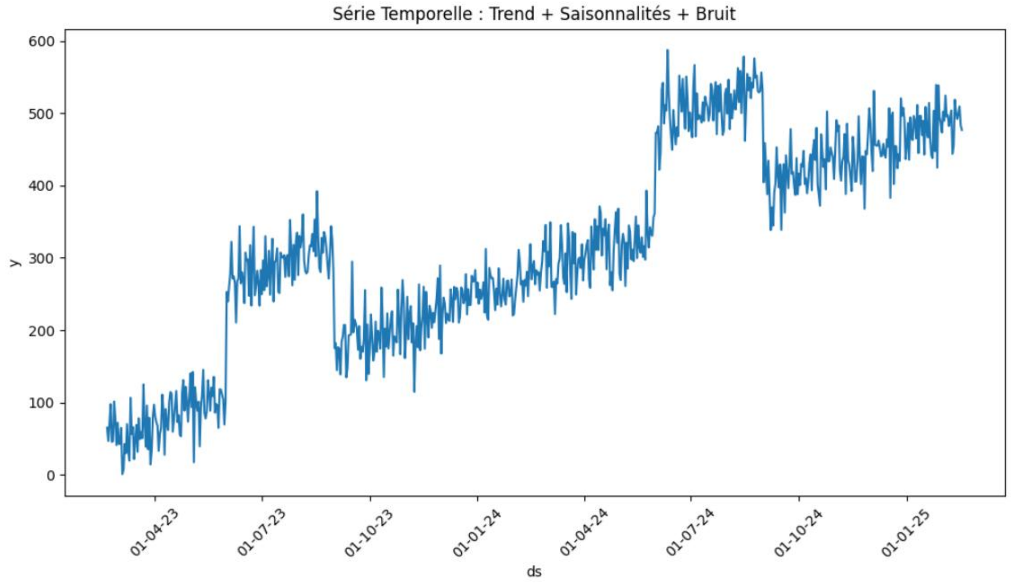

Figure 1. Small dataset (2 years)

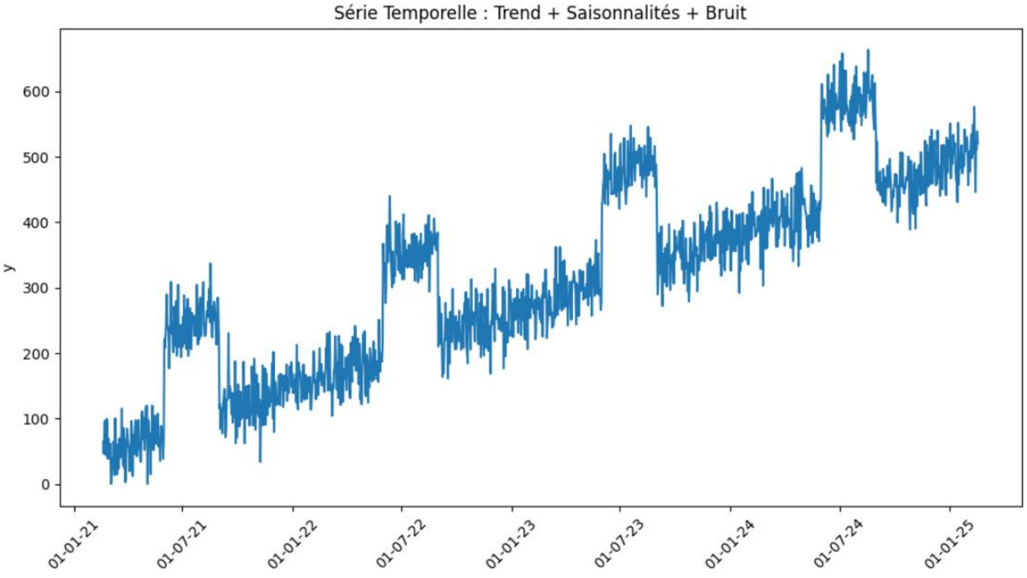

Figure 2. Average Data Set (4 years)

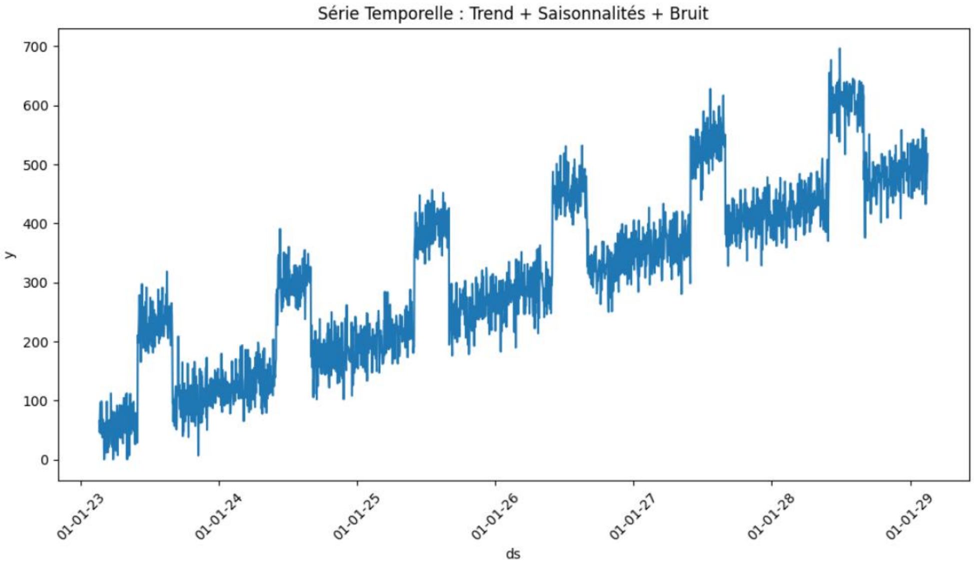

Figure 3. Large dataset (6 years)

3.2. Forecasting models

3.2.1. Prophet Model

We use the Prophet model because it can show time series data in a clear way that combines long-term trends and multiple seasonal patterns. There are two main ideas behind this method. First, it stresses clear modeling through Prophet's additive architecture, which breaks down forecasts into understandable parts, such as trend, seasonality, and external regressors. This lets practitioners directly look at and understand how each factor affects demand behavior [2]. Second, the method uses systematic optimization by doing 50 Optuna trials to carefully tune the hyperparameters [24]. The goal is to get the best possible accuracy when predicting real-world last-mile delivery data. These design choices directly address the main problems with last-mile logistics, such as dealing with seasonal demand spikes, being able to respond to changes caused by promotions or events, and the need for logistics professionals to be able to understand and make decisions. Table 1 shows the summary of proposed hyperparametric configuration and optimization for the Prophet model.

Table 1. Summary of hyperparametric configuration and optimization for Prophet model

Component	Parameter / Setting	Range / Value	Impact / Purpose
Trend	n_changepoints	10 – 30	Allows flexible modeling of structural shifts in demand
	changepoint_prior_scale	0.001 – 0.5	Controls the flexibility of trend changes
Seasonality	weekly_fourier_order	3 – 15	Sets the complexity of weekly patterns (weekday/weekend cycles)
	seasonality_prior_scale	0.1 – 20	Regulates amplitude of seasonal components
	annual seasonality	Enabled, auto regularized	Captures long-term periodic patterns while preventing overfitting
External Regressors	add_regressor()	Dynamic	Incorporates covariates (e.g., promotions, holidays, weather) into forecasts
Validation	Initial window	365 days	One year of historical data used for training
	Forecast horizon	150 days	Medium-term prediction evaluation
	Step size	90 days	Iterative shift for time-based cross-validation
	Target metric	MAE	Prioritizes operational robustness in forecasting

3.2.2. Modified Architecture of GPT-2

GPT-2, which is based on the Transformer architecture [14], uses attention mechanisms to model long-term dependencies in sequential data. Recent studies [17, 25] have shown that it can be fine-tuned to work with time series forecasting. We expanded the GPT-2 architecture, which was first made for natural language processing, to predict logistic time series. We do this by creating a new framework based on three theoretical principles. First, temporal tokenization turns digital sequences into embeddings using a 90-step sliding window method, as described in [18]. This lets the model capture medium-term contextual dependencies. Second, to avoid overfitting on small datasets (≤8,000 data points), we use complexity optimization by letting the number of layers (from 2 to 10) and the number of attention heads (2, 4, or 8) be changed. Third, knowledge integration uses nine pre-calculated covariates, such as trends, seasonalities, and trigonometric encodings, to let the model combine historical patterns with new variables. This adaptation is enhanced by meticulous hyperparameter optimization through Optuna, coupled with sophisticated regularization techniques, tackling the difficulties encountered in deploying large language model transformers within data-restricted operational logistics settings.

The modified architecture combines a projection layer (Linear + GELU) that transforms the 9 input features into 256-D embeddings. The lightweight configuration of our GPT-2 kernel (n_embd=256, n_layer=2-10, n_head=2/4/8, dropouts ∈ [0.05,0.3]) relies on efficiency principles validated by recent literature. As demonstrated by [26-27], the controlled reduction of embedding dimensions and the number of attention heads preserves representational capabilities while optimizing the computational footprint. The dropout ranges, calibrated according to [28], provide robust regularization against overfitting. This approach is in line with the recommendations of [29] for efficient transformers and is in line with lightweight temporal architectures such as the Temporal Fusion Transformer [16]. Ultimately, a final regression layer (Linear) generates the predictions. This structure preserves the ability of transformers to model long dependencies while reducing complexity for limited data. Table 2 shows the summary of proposed hyperparametric configuration and optimization for the GPT-2 model.

Table 2. Proposed fine-tuning strategy hyperparametric configuration for GPT-2

Component	Parameter / Setting	Range / Value	Impact / Purpose
Cost Function	Huber Loss δ	0.5 – 2.0	Provides robustness to outliers while balancing sensitivity and stability
Optimizer	AdamW	weight_decay = 0.005	Optimizes network weights with regularization to prevent overfitting
Learning Rate Management	LR reduction factor	0.2	Reduces learning rate by 80% after 7 epochs without improvement
	Early stopping	10 stagnant epochs	Stops training to prevent overfitting when performance plateaus
Regularization	Gradient clipping	max norm = 1.0	Limits gradient magnitude to ensure stable training
	Layer-specific dropout	Configurable	Prevents overfitting by randomly deactivating neurons during training

4. Results and Comparative Performance Assessment

4.1. Analysis of Prophet Results

The performance of the Prophet model was evaluated on its ability to forecast short and medium-term logistics demand. The results were analyzed using two key metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), measuring the absolute accuracy and the magnitude of extreme errors, respectively as shown in Figure 4-6.

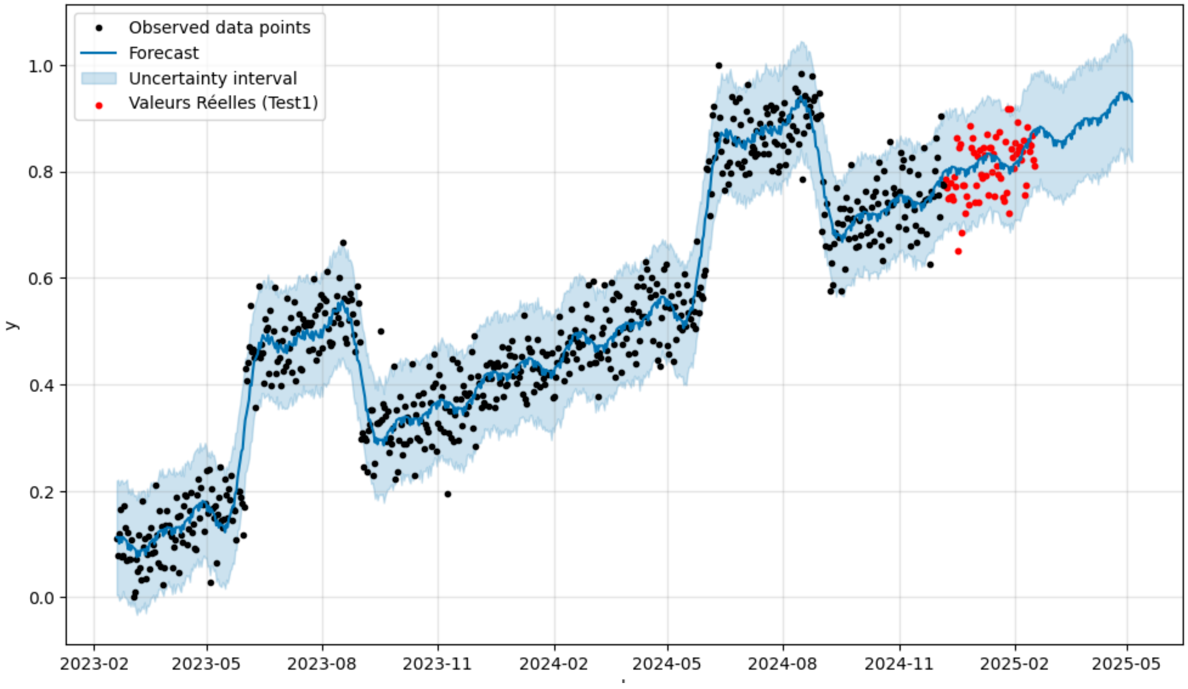

Figure 4. Predictions vs Actuals of a series of two years with prophet

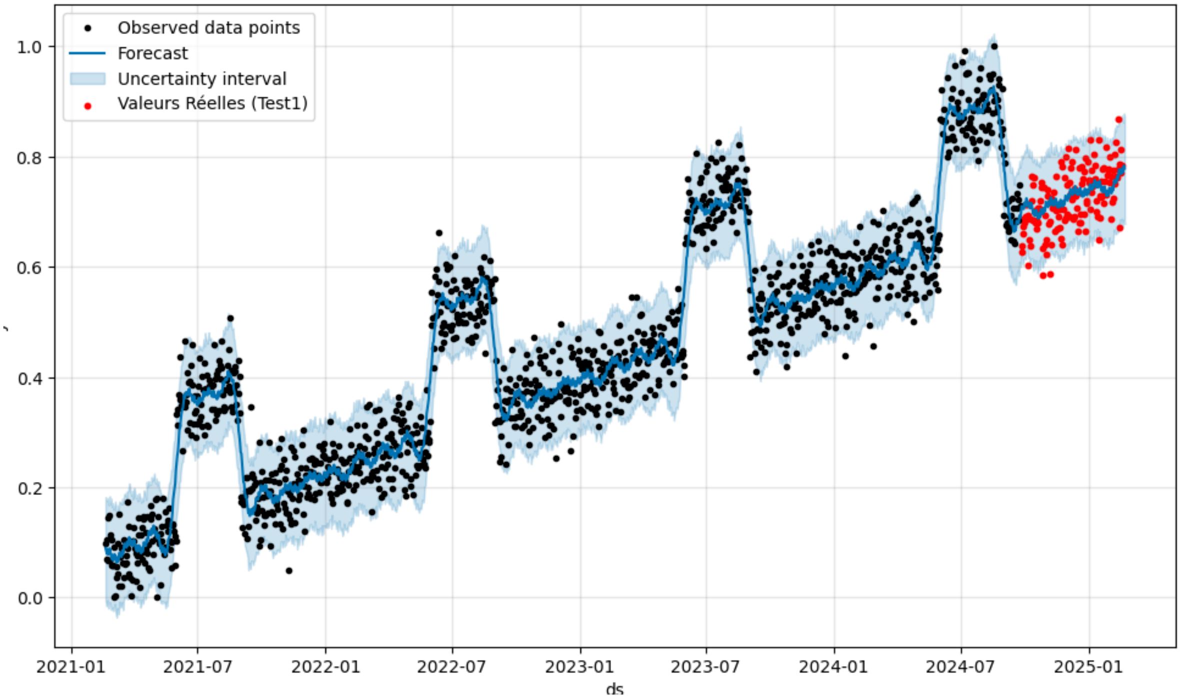

Figure 5. Predictions vs Actuals of a series of four years with prophet

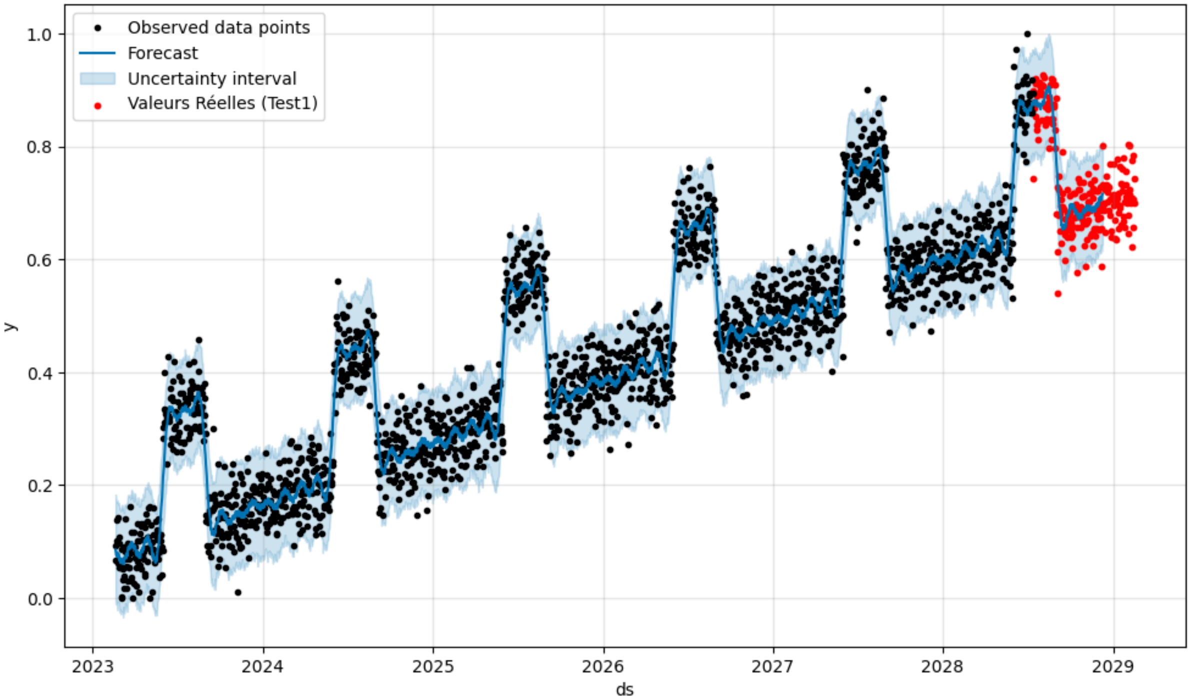

Figure 6. Predictions vs Actuals of a series of six years with prophet

4.2 Analysis of GPT-2 Results

The performance of the GPT-2 model adapted to time series was evaluated on its ability to forecast logistics demand with complex patterns. The results were analyzed via Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as show in Figure 7-8 for two-year series, Figure 9-10 for four-year series and Figure-11-12 for six-year series.

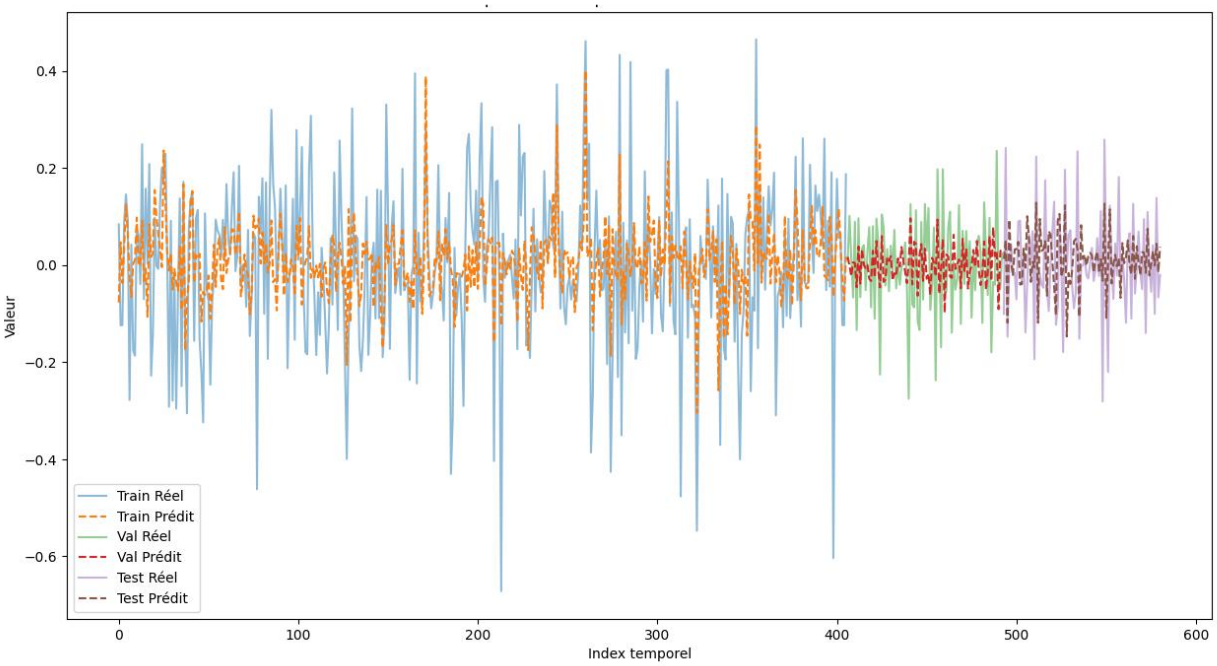

Figure 7. Comparison of different experiments training vs validation of two-year series with GPT-2

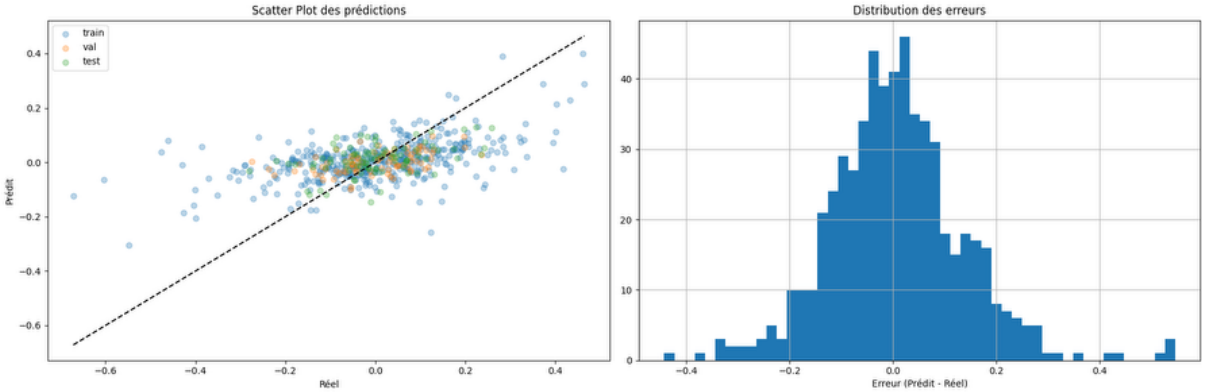

Figure 8. Scatter plot of predictions and errors of two-year series with GPT-2

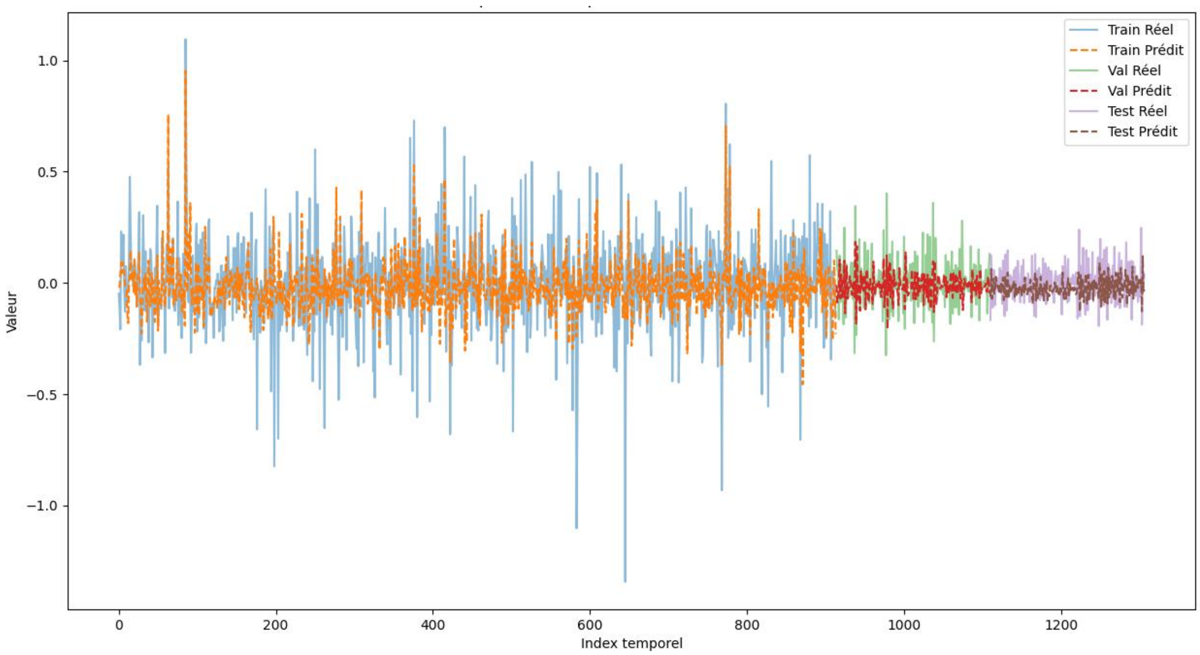

Figure 9. Comparison of different experiments training vs validation of four-year series with GPT-2

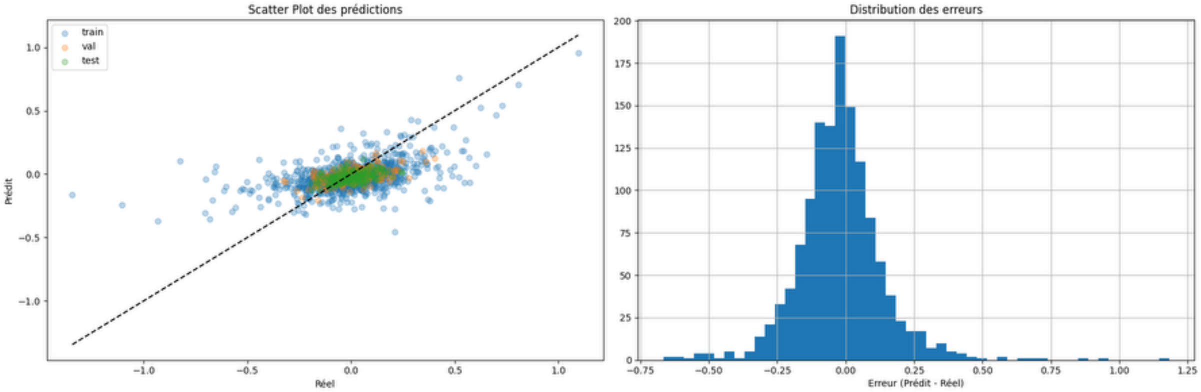

Figure 10. Scatter plot of predictions and errors of four-year series with GPT-2

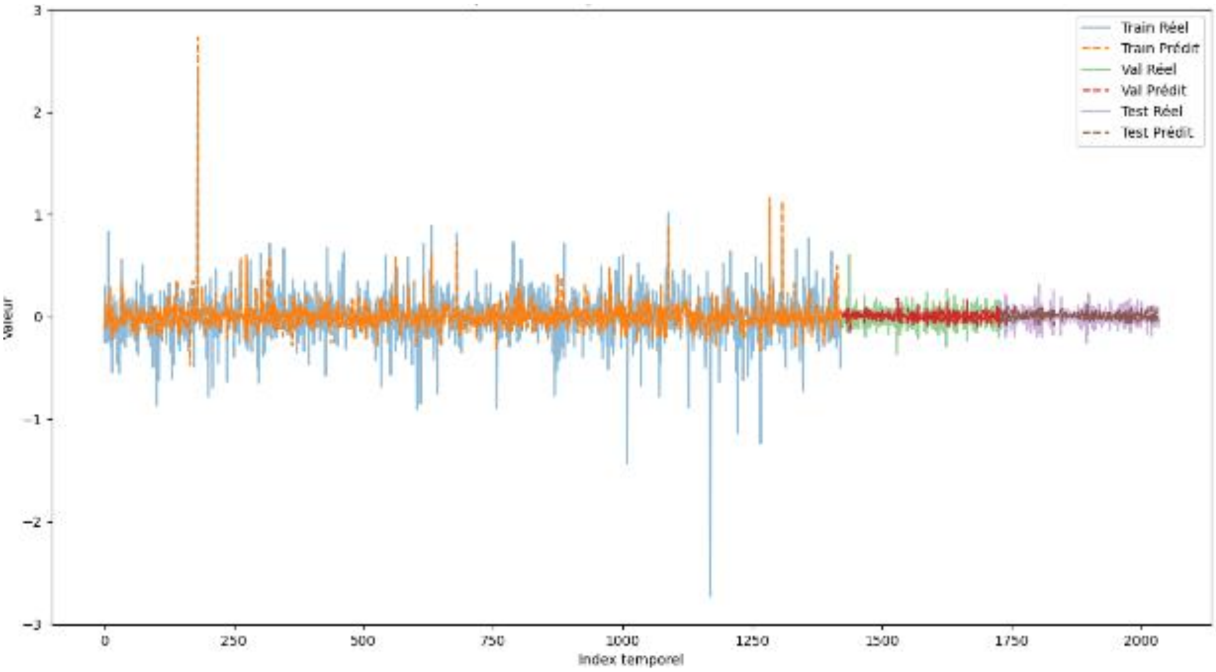

Figure 11. Comparison of different experiments training vs validation of six-year series with GPT-2

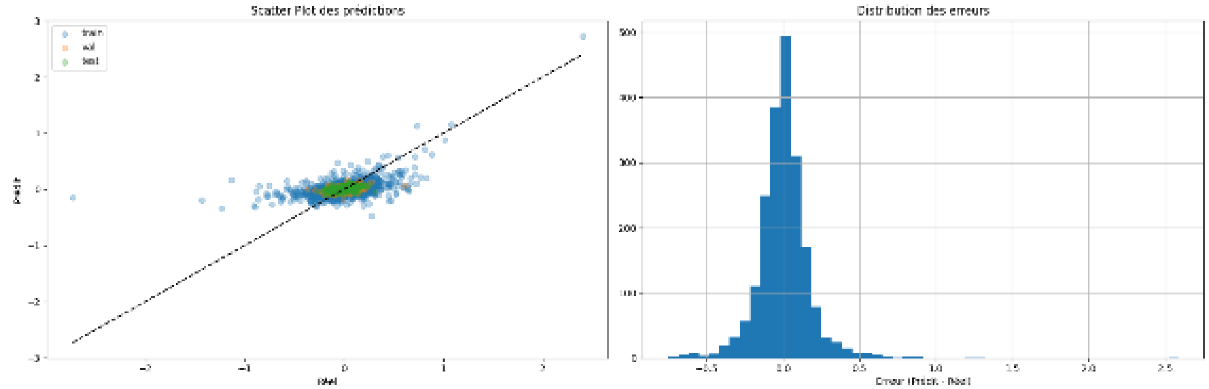

Figure 12. Scatter plot of predictions and errors of six-year series with GPT-2

The comparative research study of the Prophet and GPT-2 models reveals unique performance trends at different demand levels in last-mile logistics forecasting. Prophet consistently shows better accuracy for small, average, and large demand situations when using mean absolute error (MAE) and root mean square error (RMSE) as the main measures as shown in Table 3. For times when demand is low, Prophet has an MAE of 0.043 and an RMSE of 0.0525, which is better than GPT-2, which has an MAE of 0.0731 and an RMSE of 0.0934. In situations with medium demand, Prophet performs well, with an MAE of 0.0373 and an RMSE of 0.0466. On the other hand, GPT-2 has higher errors (MAE: 0.0616, RMSE: 0.0774), which means it is more sensitive to moderate changes. Prophet keeps making stable predictions during busy times (MAE: 0.0375, RMSE: 0.0481), but GPT-2 is still not as accurate (MAE: 0.0595, RMSE: 0.0789). These results show that GPT-2 could be useful for modeling complex dependencies. However, Prophet's clear additive decomposition and structured handling of trends and seasonality make it easier to make reliable and consistent forecasts in a variety of operational conditions.

Table 3: Comparative analysis of forecasting techniques using MAE and RMSE on different datasets

Matric	**MAE**		**RMSE**	
Model	**Prophet**	**GPT-2**	**Prophet**	**GPT-2**
Little	0.043	0.0731	0.0525	0.0934
Average	0.0373	0.616	0.0466	0.0774
Big	0.0375	0.0595	0.0481	0.0789

5. Conclusion

The proposed study compared two time series forecasting models, Prophet and GPT-2, on three datasets of different durations (2, 4, and 6 years). Each model was rigorously evaluated using metrics such as mean absolute error (MAE) and root mean square error (RMSE). The results provide a detailed comparative analysis of the performance of the two approaches. This study highlights the most suitable solutions according to use cases, while highlighting the strengths and practical applications of each model. The results summarized in the table below detail the performance of each model, to help select the most appropriate model for specific forecasting tasks. Comparative analysis (table 2) of prediction performance between Prophet (developed by Facebook) and GPT-2 reveals differences across three data classes (Small, Medium, Large). MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) metrics demonstrate the systematic superiority of Prophet across all scenarios. These results confirm the findings, which highlight Prophet's robustness to limited data. They also align with the observations, noting that specialized sequential models generally outperform generative architectures for time series. GPT-2's sensitivity to data size is particularly visible in 'Medium' category where its MAE deviates sharply (+65%) corroborates the limitations regarding out-of-distribution extrapolation.

Abbreviations

The following abbreviations are used in this manuscript:

GPT-2	Generative Pretrained Transformer
LSTM	Long Short-Term Memory
ARIMA	AutoRegressive Integrated Moving Average
SARIMA	Seasonal AutoRegressive Integrated Moving Average
HPO	Hyperparameter Optimization
LLM	Large Language Model

Acknowledgments

This work was conducted as part of the MILEX project and greatly benefited from the unwavering support of the Hassan II University of Casablanca, Ecole Centrale Casablanca, and Mohammed VI Polytechnic University. The collaborative and constructive involvement of these institutions played a pivotal role in the successful realization of this work, and we express our profound gratitude for their invaluable contributions throughout the research process. Thanks, are also due to the referees for their valuable comments.

Data Availability Statement

Not applicable.

Funding
This work was supported without any funding.

Conflicts of Interest
The author declares no conflicts of interest.

Ethical Approval and Consent to Participate
Not applicable.

References

[1]. Rao, P., Gunasekaran, A., & Goyal, S. K. (2020). Last-mile delivery: A systematic review of literature and research agenda. *Transportation Research Part E: Logistics and Transportation Review*, 144, 102098.

[2]. Taylor, S. J., & Letham, B. (2018). Forecasting at scale: The Prophet model. *The American Statistician*, 72(1), 37-45.

[3]. Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FABLE: Fast, accurate, and interpretable forecasts of multiple time series. *International Journal of Forecasting*, 36(1), 173-185.

[4]. Franceschetti, A., Honhon, D., Van Woensel, T., Bektaş, T., & Laporte, G. (2013). The trade-off between CO2 emissions and costs in vehicle routing: A case study. *Transportation Research Part D: Transport and Environment*, 18, 31-37.

[5]. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.

[6]. Nie, Y., Nguyen, N. H., Sinthupinyo, P., & Levy, C. (2023). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 39(1).

[7]. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.

[8]. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI Blog.

[9]. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

[10]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

[11]. Xu, Z., Elomri, A., Kerbache, L., & El Omri, A. (2019). The impact of weather on last-mile delivery: A literature review. *Transportation Research Part D: Transport and Environment*, 75, 115-134.

[12]. Gutierrez, A. M., & Tordecilla, R. D. (2021). Forecasting in supply chain and logistics: A review of classical and machine learning methods. *International Journal of Information Systems and Supply Chain Management*, 15(1), 1-21.

[13]. Zhang, X., Xu, J., & Ning, P. (2021). Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting. *Environmental Science and Pollution Research*, 28(41), 57560–57571.

[14]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[15]. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, J. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12), 11106–11115.

[16]. Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748–1764.

[17]. Nie, Y., Nguyen, N. H., Sinthupinyo, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. *International Conference on Learning Representations (ICLR)*.

[18]. Franceschetti, A., Bektaş, T., & Laporte, G. (2013). The impact of demand forecasting on CO2 emissions in logistics. *Transportation Research Part D: Transport and Environment*, 18, 31–37.

[19]. Chua, X. H., Li, M., & Xu, Z. (2020). Inventory optimization with hybrid models: A case study at Amazon. *INFORMS Journal on Applied Analytics*, 50(2), 116–130.

[20]. European Environment Agency. (2022). *Decarbonising transport in Europe: The role of vehicles, fuels and demand*. EEA Report.

[21]. Lundberg, O., Santén, V., & Edwards, S. (2023). *GPT-3 for sustainable logistics: A case study on emission reductions*. arXiv preprint.

[22]. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.

[23]. Wamba-Taguimdje, S. L., Fosso Wamba, S., Kala Kamdjoug, J. R., & Tchatchouang Wanko, C. E. (2020). Influence of artificial intelligence (AI) on firm performance: The role of AI-human interaction, AI-based capabilities and digital transformation. *Journal of Business Research*, 118, 50–67.

[24]. Parekh, N., Sen, A., Rajasekaran, P., Jayaseeli, J. D., & Robert, P. (2024, December). Network Intrusion Detection System Using Optuna. In *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)* (pp. 312-318). IEEE.

[25]. Kang, Y., Hyndman, R. J., & Li, F. (2021). *Feature-based forecast model selection*. arXiv preprint.

[26]. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv preprint.

[27]. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., ... & Liu, Q. (2019). *TinyBERT: Distilling BERT for natural language understanding*. arXiv preprint.

[28]. Merity, S., Keskar, N. S., & Socher, R. (2017). *Regularizing and optimizing LSTM language models*. arXiv preprint.

[29]. Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2020). Efficient transformers: A survey. *ACM Computing Surveys (CSUR)*, 55(6), 1–28.