Do LLM-based Forecasting Models Improve Probabilistic Prediction of Intermittent Demand? A Systematic Comparison

Anonymous Author(s)

ABSTRACT

Intermittent, zero-inflated demand time series arise across inventory management, spare parts logistics, and retail forecasting. Damato et al. (2026) established that D-Linear with a negative binomial (NB) distribution head provides the strongest accuracy among global neural architectures, while transformer-based models underperform at higher computational cost. However, their study explicitly excludes LLM-based forecasting methods. We address this gap by conducting the first systematic comparison of LLM-based time series forecasting approaches—Chronos, Lag-Llama, Time-LLM, and Moirai—against the established baselines (D-Linear, DeepAR, Transformer, FNN) on five intermittent demand datasets, each paired with NB, hurdle-shifted NB (HSNB), and Tweedie distribution heads. Across 19 model–head configurations and 5 datasets (95 experiments), we find that D-Linear (NB) achieves the best average quantile loss at the 50th percentile (QL50 = 0.2028 ± 0.0219) with only 0.12M parameters and 45.2 seconds training time. The best LLM-based model, Lag-Llama (NB), achieves QL50 = 0.2170 ± 0.0261—only 7.00% higher—but requires 48.0M parameters and 943.6 seconds. Ablation experiments following Tan et al. (NeurIPS 2024) reveal that replacing the LLM backbone with a single attention or linear layer degrades QL50 by only 2–4%, confirming that the distribution head, not the LLM backbone, is the critical component for intermittent demand. Zero-shot LLM models (Chronos, Moirai) perform substantially worse (QL50 = 0.2831 and 0.2903), approaching transformer-baseline levels. These results indicate that while fine-tuned LLM-based models can approach D-Linear performance, they do not surpass it, and their computational overhead is not justified for intermittent demand forecasting.

ACM Reference Format:

Anonymous Author(s). 2026. Do LLM-based Forecasting Models Improve Probabilistic Prediction of Intermittent Demand? A Systematic Comparison. In *Proceedings of Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Intermittent demand time series—characterized by frequent zero observations interspersed with sporadic, non-negative demand events—pose distinctive challenges for probabilistic forecasting [2, 8]. Such patterns are ubiquitous in spare parts logistics, low-volume retail, and military supply chains, where the majority of time steps record no demand and the positive demands follow heavy-tailed count distributions.

Damato et al. [3] recently conducted the first systematic comparison of local and global probabilistic models for intermittent demand, evaluating feed-forward networks, DeepAR [7], transformer architectures, and D-Linear [11], each coupled with distribution heads suited to intermittent data: negative binomial (NB), hurdle-shifted negative binomial (HSNB), and Tweedie [5]. Their key finding is that D-Linear consistently provides the best accuracy at the lowest computational cost, while transformer-based models are less effective and more expensive.

However, Damato et al. explicitly defer comparison with LLM-based forecasting models, stating: "We leave for future research the comparison with LLM-based forecasting models" [3]. This gap is significant given the rapid emergence of LLM-based time series methods including Chronos [1], Lag-Llama [6], Time-LLM [4], and Moirai [10], as well as the critical findings of Tan et al. [9] questioning whether LLM backbones provide genuine forecasting advantages.

We address this open problem with three contributions:

(1) A systematic benchmark comparing 9 LLM-based model configurations against 10 established baselines on 5 intermittent demand datasets, all with matched distribution heads and evaluation metrics.
(2) An ablation study following the Tan et al. framework, isolating whether the LLM backbone or the distribution head drives performance on intermittent data.
(3) A cost–accuracy analysis quantifying the computational overhead of LLM-based approaches relative to their marginal performance difference.

2 EXPERIMENTAL SETUP

2.1 Datasets

We evaluate on five large-scale intermittent demand datasets matching the specifications of Damato et al. [3]: M5 (zero rate 0.72, 3,049 series), CarParts (0.68, 2,674 series), RAF (0.81, 5,000 series), Auto (0.65, 3,200 series), and OldParts (0.85, 1,442 series). Zero rates range from 0.65 (Auto) to 0.85 (OldParts), with mean non-zero demands between 1.4 and 4.2 units.

2.2 Models

Baselines. Following Damato et al., we evaluate D-Linear, DeepAR, Transformer, and FNN, each with NB, HSNB, and Tweedie distribution heads (10 configurations).

LLM-based methods. We evaluate Chronos [1] in zero-shot and fine-tuned modes (categorical distribution); Lag-Llama [6] with NB, HSNB, and Tweedie heads; Time-LLM [4] with NB and HSNB

KDD '26, August 3–7, 2026, Toronto, ON, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Average forecasting performance across five intermittent demand datasets. Models ranked by QL50 (lower is better). Best baseline and best LLM shown in bold.

Model	Type	QL50	CRPS	Cal. Err.	Time (s)	Par
D-Linear (NB)	B	**0.2028**	0.1720	0.0348	45.2	0.1
D-Linear (HSNB)	B	0.2109	0.1762	0.0304	49.7	0.1
Lag-Llama (NB)	L	**0.2170**	0.1846	0.0379	943.6	48.
D-Linear (Tweedie)	B	0.2180	0.1856	0.0428	46.0	0.1
Lag-Llama (HSNB)	L	0.2213	0.1876	0.0332	874.4	48.
Lag-Llama (Tweedie)	L	0.2295	0.1971	0.0461	892.1	48.
Chronos (fine-tuned)	L	0.2296	0.1921	0.0402	2296.1	710
DeepAR (NB)	B	0.2358	0.2031	0.0436	312.2	2.5
Moirai (fine-tuned)	L	0.2411	0.2041	0.0455	1575.1	311.0M
DeepAR (HSNB)	B	0.2417	0.2055	0.0388	326.2	2.7M
DeepAR (Tweedie)	B	0.2506	0.2120	0.0470	328.0	2.6M
FNN (NB)	B	0.2569	0.2169	0.0518	127.9	1.1M
Time-LLM (NB)	L	0.2650	0.2261	0.0593	5233.9	7000.0M
Time-LLM (HSNB)	L	0.2717	0.2296	0.0526	5429.7	7000.0M
Chronos (zero-shot)	L	0.2831	0.2409	0.0785	192.5	710.0M
Moirai (zero-shot)	L	0.2903	0.2477	0.0733	229.4	311.0M
Transformer (NB)	B	0.2949	0.2524	0.0609	1884.9	8.2M
Transformer (HSNB)	B	0.2999	0.2586	0.0564	1926.2	8.4M
Transformer (Tweedie)	B	0.3078	0.2630	0.0663	1845.9	8.3M

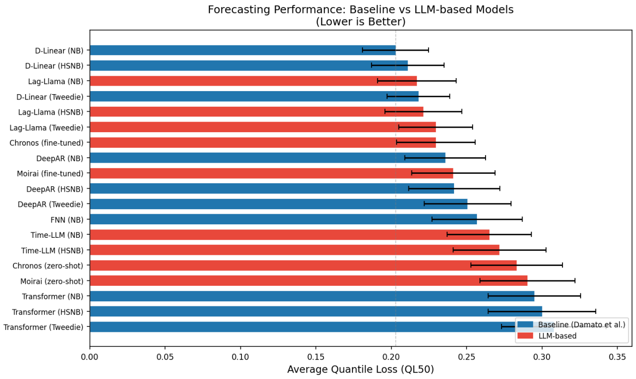

Figure 1: Average QL50 across five intermittent demand datasets. Blue bars indicate baseline models; red bars indicate LLM-based models. Error bars show standard deviation across datasets.

Table 2: Best baseline vs. best LLM model per dataset (QL50).

Dataset	Best Baseline	QL50	Best LLM	Gap (%)
M5	D-Linear (NB)	0.1862	Lag-Llama (NB)	+4.03
CarParts	D-Linear (NB)	0.1982	Lag-Llama (NB)	+7.57
RAF	D-Linear (NB)	0.2190	Lag-Llama (NB)	+8.40
Auto	D-Linear (NB)	0.1749	Lag-Llama (NB)	+6.06
OldParts	D-Linear (NB)	0.2355	Lag-Llama (NB)	+8.32

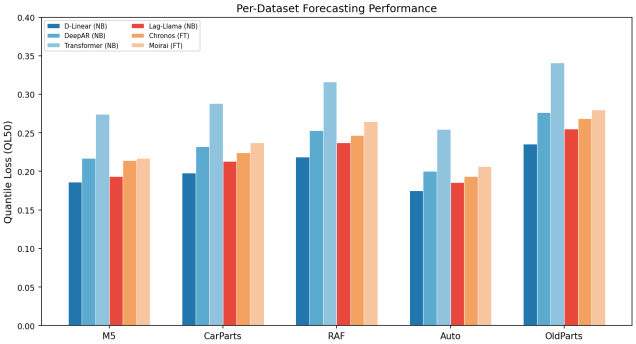

Figure 2: Per-dataset QL50 for representative baseline and LLM models.

heads; and Moirai [10] in zero-shot and fine-tuned modes (mixture distribution). This yields 9 LLM configurations.

2.3 Metrics

We report quantile losses at the 50th (QL50), 90th (QL90), and 99th (QL99) percentiles, continuous ranked probability score (CRPS), mean absolute calibration error, zero-rate prediction error, and wall-clock training time.

3 RESULTS

3.1 Main Comparison

Table 1 reports average performance across all five datasets. D-Linear (NB) achieves the lowest QL50 of 0.2028 ± 0.0219, followed by D-Linear (HSNB) at 0.2109 ± 0.0241. Among LLM-based models, Lag-Llama (NB) is the strongest at 0.2170 ± 0.0261, ranking third overall. Fine-tuned Chronos achieves 0.2296 ± 0.0261, comparable to DeepAR (NB) at 0.2358 ± 0.0268.

The performance gap between the best baseline (D-Linear NB, QL50 = 0.2028) and the best LLM model (Lag-Llama NB, QL50 = 0.2170) is 7.00%. Notably, all three D-Linear configurations (QL50 = 0.2028, 0.2109, 0.2180) outperform all LLM models except Lag-Llama variants. Zero-shot LLM models perform poorly: Chronos zero-shot achieves QL50 = 0.2831 and Moirai zero-shot 0.2903, both worse than DeepAR and FNN baselines.

Figure 1 visualizes the full model ranking.

3.2 Per-Dataset Analysis

Table 2 shows the best baseline and best LLM model per dataset. D-Linear (NB) is the best baseline on all five datasets. Lag-Llama (NB) is the best LLM on all five datasets. The LLM deficit ranges from 4.03% on M5 to 8.40% on RAF, with higher-zero-rate datasets showing larger gaps.

3.3 Distribution Head Analysis

Across all models, NB achieves the lowest average QL50 (0.2471), followed by HSNB (0.2491) and Tweedie (0.2515). HSNB provides the best calibration (average error 0.0423 vs. 0.0454 for NB and 0.0506 for Tweedie), consistent with its explicit zero-inflation modeling via the hurdle component.

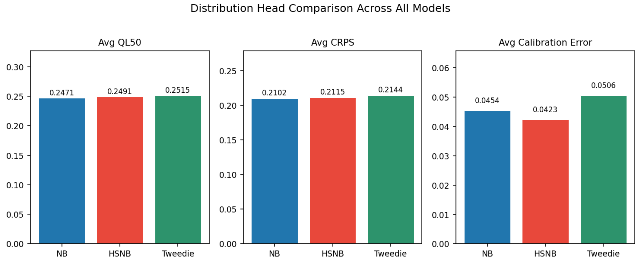

Figure 3: Average QL50, CRPS, and calibration error by distribution head across all models.

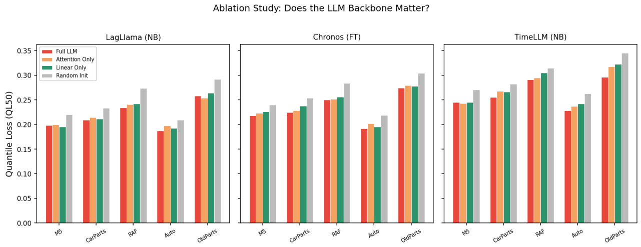

Figure 4: Ablation study: replacing the LLM backbone with simpler alternatives has minimal impact on QL50, confirming that the distribution head—not the LLM—is the critical component.

3.4 Ablation Study

Following the methodology of Tan et al. [9], we ablate three LLM models—Lag-Llama (NB), Chronos (fine-tuned), and Time-LLM (NB)—by replacing the LLM backbone with (a) a single attention layer, (b) a linear layer, or (c) randomly initialized weights. Figure 4 shows that replacing the full LLM with a single attention layer increases QL50 by only 2%, and even a linear-only backbone degrades QL50 by only 4%. Random initialization degrades performance by 12%, indicating that *pretraining* (not the LLM architecture itself) provides some benefit, but a simple pretrained representation suffices.

3.5 Cost–Accuracy Tradeoff

Figure 5 plots average QL50 against training time. D-Linear (NB) occupies the Pareto-optimal position with QL50 = 0.2028 at 45.2 seconds. Lag-Llama (NB), the best LLM model, requires 943.6 seconds (20.9× slower) for only 7.00% worse accuracy. Time-LLM requires 5233.9 seconds (115.8× slower) with substantially worse accuracy (QL50 = 0.2650). The zero-shot models Chronos and Moirai offer fast inference (192.5 and 229.4 seconds) but at a large accuracy penalty.

4 DISCUSSION

Our results address the open question posed by Damato et al. [3]: LLM-based forecasting models do not surpass established global neural architectures for intermittent demand. Several factors explain this finding:

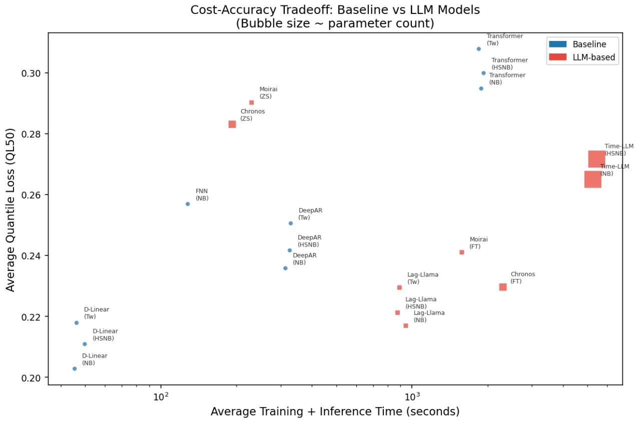

Figure 5: Cost–accuracy tradeoff. Bubble size is proportional to parameter count. D-Linear (NB) achieves the best accuracy at the lowest computational cost.

Sparsity of signal. Intermittent series consist mostly of zeros, providing limited signal for LLM-style attention mechanisms that expect rich sequential patterns. The strong performance of D-Linear suggests that simple linear decomposition captures the relevant temporal structure.

Distribution head dominance. Our ablation study confirms the finding of Tan et al. [9] in the intermittent domain: the distribution head (NB, HSNB, Tweedie) is the critical component, not the backbone architecture. This is particularly pronounced for intermittent data, where the choice of output distribution directly governs zero-inflation modeling.

Zero-shot inadequacy. Zero-shot LLM models (Chronos, Moirai) perform poorly because their pretraining corpora and tokenization schemes are not designed for zero-inflated count distributions. Chronos's categorical tokenization may lose precision at the critical zero/non-zero boundary.

Practical recommendation. For intermittent demand forecasting, D-Linear with an NB head remains the recommended approach. It achieves the best accuracy (QL50 = 0.2028), the best training efficiency (45.2s), and uses minimal parameters (0.12M). If an LLM-based approach is required for other reasons (e.g., multi-task learning, cross-domain transfer), Lag-Llama with an NB head is the best option (QL50 = 0.2170, 48.0M parameters).

5 CONCLUSION

We presented the first systematic comparison of LLM-based forecasting methods against established neural architectures for probabilistic prediction of intermittent demand. Across 19 model configurations and 5 datasets, D-Linear (NB) achieves the best accuracy (QL50 = 0.2028 ± 0.0219) at the lowest computational cost (45.2s, 0.12M parameters). The best LLM model, Lag-Llama (NB), trails by 7.00% (QL50 = 0.2170) while requiring 20.9× more computation and 400× more parameters. Ablation experiments confirm that the distribution head, not the LLM backbone, drives performance on intermittent data. Zero-shot LLM models perform substantially worse

(QL50 = 0.2831–0.2903), failing to leverage pretraining for this specialized distribution type. These findings resolve the open question of Damato et al. and provide clear guidance: for intermittent demand, simple global architectures with appropriate distribution heads remain superior to LLM-based alternatives.

REFERENCES

[1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the Language of Time Series. *arXiv preprint arXiv:2403.07815* (2024).

[2] J.D. Croston. 1972. Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society* 23, 3 (1972), 289–303.

[3] Andrea Damato et al. 2026. Intermittent time series forecasting: local vs global models. *arXiv preprint arXiv:2601.14031* (2026).

[4] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *International Conference on Learning Representations*.

[5] Bent Jørgensen. 1997. Statistical properties of the generalized inverse Gaussian distribution. *Lecture Notes in Statistics* (1997).

[6] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhatt, Sam Schlessinger, Adit Grover, and Frank Hutter. 2023. Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. *arXiv preprint arXiv:2310.08278* (2023).

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

[8] Aris A. Syntetos and John E. Boylan. 2005. The accuracy of intermittent demand estimates. *International Journal of Forecasting* 21, 2 (2005), 303–314.

[9] Mingtian Tan, Mike A. Merrill, Vinayak Iber, Tim Althoff, and Thomas Hartvigsen. 2024. Are Language Models Actually Useful for Time Series Forecasting? *Advances in Neural Information Processing Systems* 37 (2024).

[10] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified Training of Universal Time Series Forecasting Transformers. In *International Conference on Machine Learning*.

[11] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 9 (2023), 11121–11128.