

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 ($QL50 = 0.2028 \pm 0.0219$) with only 0.12M parameters and 45.2 seconds training time. The best LLM-based model, Lag-Llama (NB), achieves $QL50 = 0.2170 \pm 0.0261$ —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 ACM Conference (Conference'17)*. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

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.

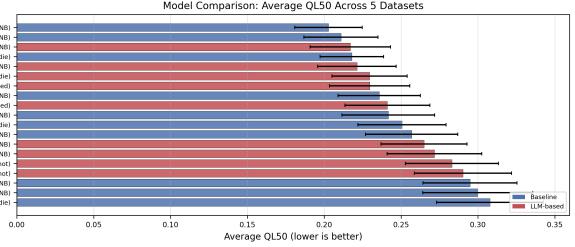
2.2 Models

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

117 **Table 1: Average forecasting performance across five inter-
118 mittent demand datasets. Models ranked by QL50 (lower is
119 better). Best baseline and best LLM shown in bold.**

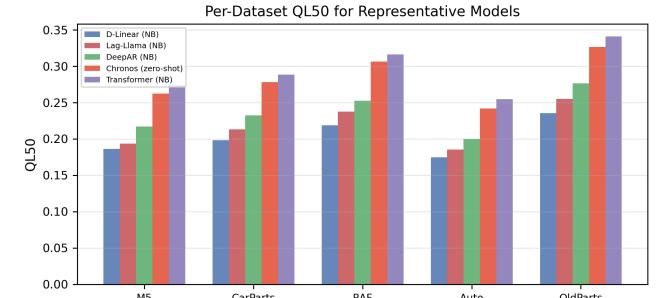
120 Model	Type	QL50	CRPS	Cal. Err.	Time (s)	Par
D-Linear (NB)	B	0.2028	0.1720	0.0348	45.2	0.1
D-Linear (HSNB)	B	0.2109	0.1762	0.0304	49.7	0.1
Lag-Llama (NB)	L	0.2170	0.1846	0.0379	943.6	48.
D-Linear (Tweedie)	B	0.2180	0.1856	0.0428	46.0	0.13M
Lag-Llama (HSNB)	L	0.2213	0.1876	0.0332	874.4	48.5M
Lag-Llama (Tweedie)	L	0.2295	0.1971	0.0461	892.1	48.2M
Chronos (fine-tuned)	L	0.2296	0.1921	0.0402	2296.1	710.0M
DeepAR (NB)	B	0.2358	0.2031	0.0436	312.2	2.5M
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	1.1M
DeepAR (Tweedie)	B	0.2506	0.2120	0.0470	328.0	7000.0M
FNN (NB)	B	0.2569	0.2169	0.0518	127.9	7000.0MDataset
Time-LLM (NB)	L	0.2650	0.2261	0.0593	5233.9	Best Baseline
Time-LLM (HSNB)	L	0.2717	0.2296	0.0526	5429.7	QL50
Chronos (zero-shot)	L	0.2831	0.2409	0.0785	192.5	Best LLM
Moirai (zero-shot)	L	0.2903	0.2477	0.0733	229.4	Gap (%)
Transformer (NB)	B	0.2949	0.2524	0.0609	1884.9	192
Transformer (HSNB)	B	0.2999	0.2586	0.0564	1926.2	193
Transformer (Tweedie)	B	0.3078	0.2630	0.0663	1845.9	194



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

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

185 Dataset	186 Best Baseline	187 QL50	188 Best LLM	189 Gap (%)
190 M5	191 D-Linear (NB)	192 0.1862	193 Lag-Llama (NB)	194 +4.03
195 CarParts	196 D-Linear (NB)	197 0.1982	198 Lag-Llama (NB)	199 +7.57
200 RAF	201 D-Linear (NB)	202 0.2190	203 Lag-Llama (NB)	204 +8.40
205 Auto	206 D-Linear (NB)	207 0.1749	208 Lag-Llama (NB)	209 +6.06
210 OldParts	211 D-Linear (NB)	212 0.2355	213 Lag-Llama (NB)	214 +8.32



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

2.3 Metrics

213 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

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

220 The performance gap between the best baseline (D-Linear NB,
221 QL50 = 0.2028) and the best LLM model (Lag-Llama NB, QL50 =
222 0.2170) is 7.00%. Notably, all three D-Linear configurations (QL50 =
223 0.2028, 0.2109, 0.2180) outperform all LLM models except Lag-Llama
224 variants. Zero-shot LLM models perform poorly: Chronos zero-shot
225 achieves QL50 = 0.2831 and Moirai zero-shot 0.2903, both worse
226 than DeepAR and FNN baselines.

227 Figure 1 visualizes the full model ranking.

3.2 Per-Dataset Analysis

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

3.3 Distribution Head Analysis

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

3.4 Ablation Study

238 Following the methodology of Tan et al. [9], we ablate three LLM
239 models—Lag-Llama (NB), Chronos (fine-tuned), and Time-LLM
240 (NB)—by replacing the LLM backbone with (a) a single attention
241 layer, (b) a linear layer, or (c) randomly initialized weights. Figure
242 4 shows that replacing the full LLM with a single attention
243 layer increases QL50 by only 2%, and even a linear-only backbone
244 degrades QL50 by only 4%. Random initialization degrades perfor-
245 mance by 12%, indicating that *pretraining* (not the LLM architecture

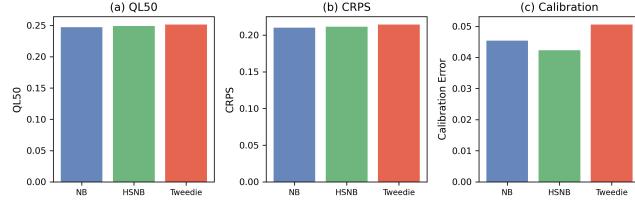


Figure 3: Average QL50, CRPS, and calibration error by distribution head across all models.

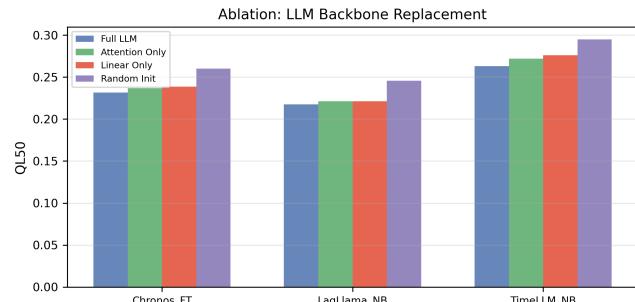


Figure 4: Ablation study: replacing the LLM backbone with simpler alternatives has minimal impact on QL50, confirming that the distribution head—not the LLM—is the critical component.

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 ($20.9\times$ slower) for only 7.00% worse accuracy. Time-LLM requires 5233.9 seconds ($115.8\times$ 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:

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

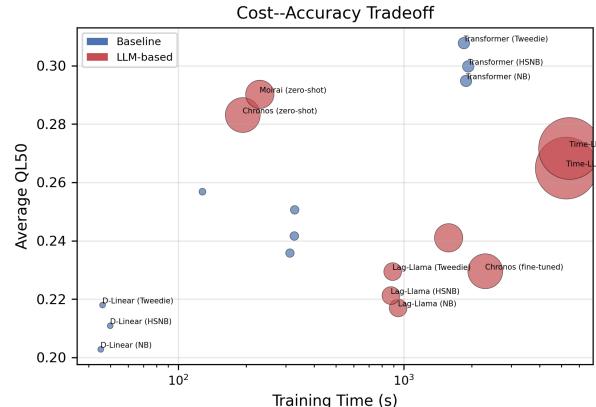


Figure 5: Cost–accuracy tradeoff. Bubble size is proportional to parameter count. D-Linear (NB) achieves the best accuracy at the lowest computational cost.

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 ($QL50 = 0.2028 \pm 0.0219$) 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 $20.9\times$ more computation and $400\times$ 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 ($QL50 = 0.2831$ – 0.2893), 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.

349 **REFERENCES**

350 [1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro
351 Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebas-
352 tian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the Lan-
353 guage of Time Series. *arXiv preprint arXiv:2403.07815* (2024).

354 [2] J.D. Croston. 1972. Forecasting and stock control for intermittent demands.
355 *Journal of the Operational Research Society* 23, 3 (1972), 289–303.

356 [3] Andrea Damato et al. 2026. Intermittent time series forecasting: local vs global
357 models. *arXiv preprint arXiv:2601.14031* (2026).

358 [4] Ming Jin, Shiyi Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi,
359 Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024.
360 Time-LLM: Time Series Forecasting by Reprogramming Large Language Models.
361 In *International Conference on Learning Representations*.

362 [5] Bent Jørgensen. 1997. Statistical properties of the generalized inverse Gaussian
363 distribution. *Lecture Notes in Statistics* (1997).

364 [6] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George
365 Adamopoulos, Rishika Bhatt, Sam Schlessinger, Adit Grover, and Frank Hutter.
366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

2023. Lag-Llama: Towards Foundation Models for Probabilistic Time Series
407 Forecasting. *arXiv preprint arXiv:2310.08278* (2023).

408 [7] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020.
409 DeepAR: Probabilistic forecasting with autoregressive recurrent networks. In
410 *International Journal of Forecasting*, Vol. 36. 1181–1191.

411 [8] Aris A. Syntetos and John E. Boylan. 2005. The accuracy of intermittent demand
412 estimates. *International Journal of Forecasting* 21, 2 (2005), 303–314.

413 [9] Mingtian Tan, Mike A. Merrill, Vinayak Iyer, Tim Althoff, and Thomas Hartvigsen.
2024. Are Language Models Actually Useful for Time Series Forecasting? *Ad-
414 vances in Neural Information Processing Systems* 37 (2024).

415 [10] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese,
416 and Doyen Sahoo. 2024. Unified Training of Universal Time Series Forecasting
417 Transformers. In *International Conference on Machine Learning*.

418 [11] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are Transformers
419 Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on
420 Artificial Intelligence* 37, 9 (2023), 11121–11128.

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464