The full version of the paper is available at https://arxiv.org/abs/2508.09904

Beyond Naïve Prompting: Strategies for Improved Zero-shot Context-aided Forecasting with LLMs

*Arjun Ashok[1,2,3], Andrew R. Williams[1,2,3], Vincent Zhihao Zheng[1,4], Irina Rish[2,3], Nicolas Chapados[1,2,5], Étienne Marcotte[1], Valentina Zantedeschi[1,6], Alexandre Drouin[1,2,6]
[1]ServiceNow Research, [2]Mila - Québec AI Institute, [3]Université de Montréal, [4]McGill University
[5]Polytechnique Montréal, [6]Université Laval

Abstract

Forecasting in real-world settings requires models to integrate not only historical data but also relevant contextual information, often available in textual form. While recent work has shown that large language models (LLMs) can be effective context-aided forecasters via naïve direct prompting, their full potential remains underexplored. We address this gap with 4 strategies, providing new insights into the zero-shot capabilities of LLMs in this setting. ReDP improves interpretability by eliciting explicit reasoning traces, allowing us to assess the model's reasoning over the context independently from its forecast accuracy. CorDP leverages LLMs solely to refine existing forecasts with context, enhancing their applicability in real-world forecasting pipelines. IC-DP proposes embedding historical examples of context-aided forecasting tasks in the prompt, substantially improving accuracy even for the largest models. Finally, RouteDP optimizes resource efficiency by using LLMs to estimate task difficulty, and routing the most challenging tasks to larger models. Evaluated on different kinds of context-aided forecasting tasks from the CiK benchmark, our strategies demonstrate distinct benefits over naïve prompting across LLMs of different sizes and families. These results open the door to further simple yet effective improvements in LLM-based context-aided forecasting.

1 Introduction

Probabilistic time series forecasting is essential for optimal decision-making, involving predicting the evolution of various quantities over time, as well as estimating the likelihood of various scenarios [20, 35]. While research has largely focused on numerical historical observations and engineered covariates, real-world forecasts rely not only on them but also on contextual information about the problem or task in hand [20]. This has led to the emergence of a new, multimodal problem setting of *context-aided forecasting* [22, 28, 48].

Several methods have been proposed for context-aided forecasting [53], broadly classified into those that rely on training models on specific context-aided forecasting tasks [22, 54, 51, 9, 44, 28, 56] and those that use LLMs zero-shot [32, 17, 38, 48]. Among the zero-shot methods, only simple strategies have been explored, such as direct prompting [48] and autoregressive LLM processes [38] among others. These methods involve simply feeding historical numerical data and textual context into the LLM and generating forecasts timestep-by-timestep. The potential for sophisticated strategies to enhance forecast accuracy, efficiency, and interpretability remains largely unexplored. In this work, we systematically investigate 4 strategies that address different aspects of zero-shot forecasting with LLMs (illustrated in Figure 1).

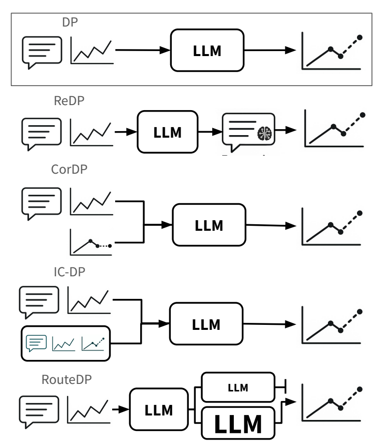


Figure 1: Direct Prompt (DP) [48] prompts the LLM with the context and historical data. Our work explores nuanced strategies with distinct benefits: ReDP improves interpretability, CorDP bootstraps LLMs on prior forecasts, IC-DP boosts performance through exemplars, and RouteDP enables accurate forecasting under resource constraints.

*Correspondence to: arjun.ashok.psg@gmail.com
NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models (BERT^{2}S).

- **ReDP: Direct Prompting with Reasoning over Context** (Section 3) improves interpretability by prompting models to output explicit context reasoning traces and comparing them with gold standard reasoning traces. This helps uncover a key failure mode of models: inability to apply their reasoning on their forecasts.
- **CorDP: Direct Prompting for Forecast Correction** (Section 4) utilizes LLMs to solely modify existing probabilistic forecasts with context, instead of forecasting from scratch. This improves models by up to 50%, and allows for practical adoption of LLMs in existing forecasting workflows.
- **IC-DP: In-Context Direct Prompting** (Section 5) explores prompting LLMs with historical examples of context-aided forecasting tasks, substantially improving accuracy even for large models.
- **RouteDP: Direct Prompting with Model Routing** (Section 6) enables accurate forecasting under resource constraints by using a small model for easy tasks and delegating difficult ones to a larger model, guided by a router. We observe substantial improvements in forecast accuracy at a fraction of the cost.

We evaluate these strategies on diverse zero-shot context-aided forecasting tasks from the Context-Is-Key (CiK) benchmark [48]. We show that such strategies can prove extremely effective in studying various capabilities of models and obtaining significantly better context-aided forecasts from LLMs across different sizes and families [52, 16].

2 Background and Experimental Protocol

The goal of context-aided forecasting is to produce statistical forecasts by incorporating relevant side information (i.e. context) [48, 44, 23, 53]. Williams et al. [48] introduce Direct Prompt (DP), which instructs LLMs to generate structured forecasts given history and context. With DP, they show that instruction-tuned LLMs can improve their performance significantly with context [48, 24], providing advantages over quantitative methods that cannot use context [23, 53]. Our work pushes DP limits with 4 strategies offering distinct advantages and considerably better performance. A discussion of related work can be found in Appendix A.

We use the Context-Is-Key (CiK) benchmark [48] to evaluate zero-shot forecasting methods. CiK contains 71 manually designed context-aided forecasting tasks from 2644 time series across 7 domains: Climatology [40], Economics [42], Energy [14], Mechanics [11], Public Safety [43], Transportation [6], and Retail [14]. CiK is uniquely suitable as accurate forecasts cannot be achieved without incorporating context [53], distinguishing it from other benchmarks [32, 28, 44, 45]. We use the Region-of-interest CRPS (RCRPS) metric [48] and experiment with Qwen-2.5-0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B [52] and Llama-3.2-1B, 3B, Llama-3-8B, Llama-3.3-70B, Llama-3.1-405B [16] models. Additional details on the experimental protocol are provided in Appendix B, and details on the implementation of each model are provided in App. H.

3 ReDP: Direct Prompting with Reasoning over Context

Context-aided forecasting requires models to correctly reason about how context should influence the forecast, and then translate this reasoning into accurate forecasts. Current evaluation approaches treat models as black boxes, focusing solely on final forecasting accuracy. This creates a critical diagnostic gap: when models fail, we cannot determine whether the failure stems from incorrect reasoning over the context, inability to apply correct reasoning to forecasting, or both. We propose ReDP, which instructs LLMs to produce explicit reasoning traces before forecasts (see appendix D.1 for the prompt), building on chain-of-thought prompting [46, 26, 10]. We evaluate reasoning correctness using gold standard traces which we generate with GPT-4.1 [1] and manually verify. We then measure correctness in reasoning by comparing reasoning traces against gold standards using GPT-4.1 as an LLM judge [18]. To understand whether correct reasoning translates to improved forecasting performance, we also measure the relative improvement in CRPS when context is provided versus when it is not, using a 50% improvement threshold to identify meaningful improvements with context. Prompts used for the judge, gold standards, and example reasoning traces are in Appendices D.2, D.4 and D.5.

The results of the reasoning quality analysis are in Table 1. First, we find that the percentage of tasks in which the model's reasoning is correct improves with the model size, across both families, and as does the percentage of tasks with a meaningful improvement with context. Next, we find that the smaller models (3B-10B) can reason correctly in only a portion of the tasks (30%-70%), and further, successfully apply it to improve with context in only a fraction of the tasks (some even failing to apply even in a single task). Next, mid-sized models (Qwen 14B, 32B) and large models (>32B) can reason correctly in almost all the tasks, and apply their reasoning correctly in about 70% of the tasks. Finally, as seen in the last two columns, in the absence of correct reasoning, tasks rarely or never see a success in improvement with context, indicating that the reasoning trace is a faithful reflection of the model's forecasting process with context. Examples from the analyses showcasing their evaluated reasoning

Model	Correct Reasoning	Improvement with Context	Correct Reasoning and Improvement with Context	Correct Reasoning but no Improvement with Context	Wrong Reasoning but Improvement with Context	Wrong Reasoning and no Improvement with Context
Llama-3.2-3B-Inst	38.9%	33.3%	16.7%	22.2%	16.7%	44.4%
Llama-3.1-8B-Inst	70.0%	30.0%	30.0%	40.0%	0.0%	30.0%
Llama-3.3-70B-Inst	100.0%	78.9%	78.9%	21.1%	0.0%	0.0%
Llama-3.1-405B-Inst	90.0%	80.0%	70.0%	20.0%	10.0%	0.0%
Qwen-2.5-3B-Inst	46.7%	0.0%	0.0%	46.7%	0.0%	53.3%
Qwen-2.5-7B-Inst	84.2%	42.1%	36.8%	47.4%	5.3%	10.5%
Qwen-2.5-14B-Inst	95.0%	80.0%	75.0%	20.0%	5.0%	0.0%
Qwen-2.5-32B-Inst	94.7%	68.4%	68.4%	26.3%	0.0%	5.3%
Qwen-2.5-72B-Inst	94.7%	78.9%	73.7%	21.1%	5.3%	0.0%

Table 1: Results of the reasoning quality analysis using ReDP. The first two columns show the percentage of tasks where models produce correct reasoning traces and achieve meaningful improvements with context respectively. The remaining columns show the joint distribution: correct reasoning with/without improvement, and incorrect reasoning with/without improvement. We find that smaller models (<10B) often reason correctly but fail to apply their reasoning, while larger models (>70B) achieve both correct reasoning and successful application.

Model	Direct Prompt (DP)	Median Corrector (Median-CorDP)			SampleWise Corrector (SampleWise-CorDP)		
		Lag-Llama	Chronos Large	arima	Lag-Llama	Chronos Large	arima
Llama3.2-1B-Inst	**0.396 ± 0.027**	**0.394 ± 0.004**	0.515 ± 0.007	0.612 ± 0.018	0.541 ± 0.009	0.634 ± 0.005	0.672 ± 0.015
Llama3.2-3B-Inst	0.687 ± 0.025	**0.344 ± 0.011**	0.455 ± 0.009	0.573 ± 0.022	0.509 ± 0.026	0.423 ± 0.007	0.663 ± 0.031
Llama3-8B-Inst	0.543 ± 0.026	**0.315 ± 0.004**	0.453 ± 0.005	0.571 ± 0.004	0.426 ± 0.009	0.410 ± 0.004	0.636 ± 0.010
Llama3.3-70B-Inst	0.230 ± 0.006	0.281 ± 0.002	0.251 ± 0.004	0.352 ± 0.006	0.223 ± 0.004	**0.215 ± 0.004**	0.311 ± 0.007
Llama3.1-405B-Inst	**0.173 ± 0.003**	0.278 ± 0.009	0.226 ± 0.004	0.257 ± 0.008	0.199 ± 0.006	0.194 ± 0.004	0.229 ± 0.008
Qwen2.5-0.5B-Inst	0.592 ± 0.027	0.633 ± 0.002	0.801 ± 0.003	0.761 ± 0.054	**0.494 ± 0.008**	0.644 ± 0.076	0.655 ± 0.055
Qwen2.5-1.5B-Inst	0.616 ± 0.018	**0.426 ± 0.013**	0.537 ± 0.003	0.682 ± 0.006	0.522 ± 0.018	0.474 ± 0.005	0.719 ± 0.013
Qwen2.5-3B-Inst	0.424 ± 0.017	0.490 ± 0.005	0.491 ± 0.004	0.597 ± 0.009	**0.398 ± 0.028**	0.451 ± 0.005	0.512 ± 0.032
Qwen2.5-7B-Inst	0.401 ± 0.006	0.419 ± 0.004	0.641 ± 0.008	0.633 ± 0.008	**0.382 ± 0.007**	0.402 ± 0.020	0.540 ± 0.011
Qwen2.5-14B-Inst	**0.247 ± 0.006**	0.315 ± 0.003	0.334 ± 0.006	0.423 ± 0.004	0.364 ± 0.006	0.410 ± 0.006	0.471 ± 0.009
Qwen2.5-32B-Inst	0.397 ± 0.008	**0.248 ± 0.004**	0.272 ± 0.005	0.329 ± 0.008	0.310 ± 0.005	0.338 ± 0.007	0.414 ± 0.009
Qwen-2.5-72B-Inst	**0.202 ± 0.009**	0.319 ± 0.008	0.358 ± 0.010	0.428 ± 0.009	0.255 ± 0.010	0.322 ± 0.010	0.386 ± 0.010
Base Quantitative Forecaster	-	0.382 ± 0.011	0.492 ± 0.004	0.636 ± 0.014	0.382 ± 0.011	0.492 ± 0.004	0.636 ± 0.014

Table 2: Aggregate results of CorDP methods on CiK, accompanied by standard errors. The best performing method for each model is in **bold**. Results on various groups of tasks are in Appendix E.2.

correctness and improvement with context are provided in Appendix D.6. These findings suggest that enhancing models's ability to effectively leverage their reasoning and apply it is a promising and tractable direction for future research.

4 CorDP - Direct Prompting for Forecast Correction

Real-world forecasting applications require both high accuracy in quantitative predictions and the ability to incorporate contextual information when available. However, current approaches attempt to replace specialized, highly-tuned quantitative models entirely with LLMs [48, 38], leading to suboptimal performance. We propose Direct Prompting for Forecast Correction (CorDP), where an LLM acts as a *forecast corrector* that bootstraps off probabilistic forecasts from quantitative models and corrects them based on context (prompt in Appendix E.1). We propose two variations: SampleWise-CorDP corrects each individual sample from the quantitative model's probabilistic forecasts, while Median-CorDP corrects the median of the base forecast multiple times. This mimics human judgmental correction [20] and also enables easy integration with existing forecasting pipelines, as it only requires adding a correction step without modifying the core quantitative forecasting infrastructure.

Table 2 presents results aggregated across all tasks from CiK, where CorDP methods achieve the best performance across 8/12 LLMs with improvements of up to 50%. Performance varies significantly with the quantitative forecaster used, with models using Lag-Llama dominating due to its superior base performance. Most models improve over their base forecaster when using CorDP, though smallest models sometimes deteriorate base forecasts while still improving over Direct Prompting. Analyses on different kinds of tasks (results in Appendix E.2) reveals SampleWise-CorDP has an advantage on tasks with a partial RoI (region of interest, the context-sensitive region within the prediction window), while Median-CorDP dominates on tasks where context influences the entire forecast shape and tasks with constraints. Example forecasts of CorDP are provided in Appendix E.3. Future work could explore fine-tuning LLMs to better match the forecast distributions of specific quantitative forecasters, and in analyzing the distributional properties of LLM-generated versus quantitative forecasts.

5 IC-DP - In-Context Direct Prompting

Real-world forecasting often involves recurring domain-specific contexts such as seasonal heat waves in electricity demand [44, 45]. Training models to handle these contexts requires costly model selection, training, and maintenance, while LLMs can instead leverage in-context learning for zero-shot forecasting [5]. To evaluate this capability, we propose In-Context Direct Prompt (IC-DP), which extends Direct Prompt by including example context-aided forecasting tasks with histories, contexts, and ground truths. We evaluate IC-DP using a single past instance of the same context-aided forecasting task as the example, where the time series and textual context differ but the context's influence on the prediction window is the same (see Appendix F.4 for examples).

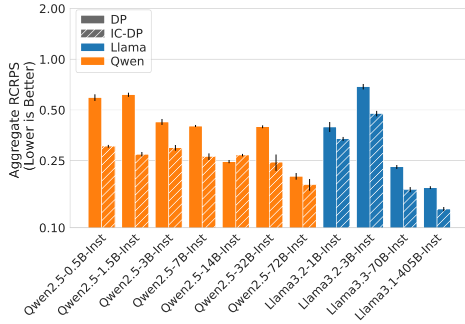


Figure 2: Aggregate results of models with DP and IC-DP respectively, accompanied by standard errors. IC-DP significantly improves the performance of 10/11 tested models.

Aggregate results in Figure 2 show that IC-DP improves 10 of 11 models, with gains of 14–55.7% for small models, 20–40% for mid- and large models, and even a 25% boost for Llama-405B-Inst, unlike CorDP which mainly benefits smaller models. Inspecting the performance of IC-DP on various kinds of tasks (Table 11) IC-DP yields the largest gains on tasks where the forecast is fully shaped by context (full-ROI), with strong improvements in ROI and constraint RCRPS. Qwen-2.5-14B remains an outlier, degrading by 9% on average, consistent with its strong zero-shot capabilities where modifications can be harmful. Future work could further consider using synthetic in-context examples to broaden applicability. Forecasts of models with IC-DP with the respective in-context examples used are provided in Appendix F.4.

6 RouteDP: Direct Prompt with Model Routing

Larger LLMs perform better at context-aided forecasting but are expensive for real-world deployment. Model routing strategies [34, 31] can be used to allocate resources adaptively, sending challenging cases to large models while routing easier tasks to smaller ones. We propose RouteDP, with which a *main model* (e.g., Qwen 0.5B) handles most tasks, while a *large model* (Llama-3.1-405B-Inst) handles the most difficult ones. A *router model* assigns difficulty scores (0-1) to tasks based on context and history. For a compute budget, the k most difficult tasks go to the large model, remaining $N-k$ tasks to the main model. We compare to random and ideal routing baselines, testing Qwen family models as routers and main models, with Llama-3.1-405B-Inst as the large model. We vary k from 0 to 71 tasks to measure performance as a function of routing budget. Additional details on the protocol are provided in Appendix G.

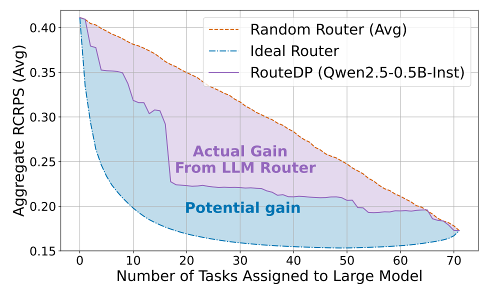


Figure 3: The plot shows the average RCRPS achieved using Qwen2.5-0.5B-Inst as the main model as an increasing % of tasks are routed to the large model (Llama-405B-Inst), using Qwen2.5-0.5B-Inst as the router model, compared to random and ideal routers. The router captures 66% of the possible area between random and ideal routing. Area captured by other models are shown in Appendix G.3.

RouteDP achieves significantly better performance than random routing (see Figure 3), with Qwen2.5-0.5B-Inst capturing 66% of the area between random and ideal routing. Routing <20% of tasks yields 46.6% improvement (see Table 12), with most gains achieved at 20-40% routing and diminishing returns beyond. Small models benefit the most: Qwen2.5-0.5B-Inst achieves 46.6% improvement vs 16.6% for Qwen2.5-14B-Inst, reflecting the larger potential improvement for smaller models. Each model acts as its own best router, with small models being disproportionately effective routers. RouteDP enables better forecasting performance as compute budget increases, offering immediate practical benefits for real-world deployment.

7 Conclusion

This work builds on recent research on using large language models (LLMs) for zero-shot, context-aided time series forecasting. We introduce four complementary strategies , namely Direct Prompting with Reasoning over Context (ReDP), Direct Prompting for Forecast Correction (CorDP), In-Context Direct Prompting (IC-DP), and Direct Prompting with Model Routing (RouteDP) which offer different benefits over naïve direct prompting (DP), improving interpretability, performance and resource efficiency. We recommend practitioners use: (i) ReDP when seeking to understand how models interpret context and analyze systematic errors, (ii) CorDP to add context-aided forecasting capabilities to custom-engineered forecasting pipelines, (iii) IC-DP when prior historical examples are available, and (iv) RouteDP to balance accuracy and computational cost. We discuss future work in Appendix I.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] 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. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[3] Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Nicolas Chapados, and Alexandre Drouin. TACTiS-2: Better, faster, simpler attentional copulas for multivariate time series. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=xtOydkE1Ku`.

[4] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, fifth edition, 2015.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Chao Chen, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. Freeway performance measurement system: mining loop detector data. *Transportation research record*, 1748(1):96–102, 2001.

[7] Yixin Dong, Charlie F Ruan, Yaxing Cai, Ziyi Xu, Yilong Zhao, Ruihang Lai, and Tianqi Chen. Xgrammar: Flexible and efficient structured generation engine for large language models. In *Eighth Conference on Machine Learning and Systems*.

[8] Alexandre Drouin, Étienne Marcotte, and Nicolas Chapados. Tactis: Transformer-attentional copulas for time series. In *International Conference on Machine Learning*, pages 5447–5493. PMLR, 2022.

[9] Patrick Emami, Zhaonan Li, Saumya Sinha, and Truc Nguyen. Syscaps: Language interfaces for simulation surrogates of complex systems. *arXiv preprint arXiv:2405.19653*, 2024.

[10] Stephanie Fu, Sushant Arora, Benjamin Heinzerling, Xiang Lisa Lee, Benjamin Peters, Qian Wang, John Thickstun, Elizabeth Dyer, and Dzmitry Bahdanau. Chain-of-thought hub: A crowdsourced benchmark for evaluating reasoning. *arXiv preprint arXiv:2306.03314*, 2023.

[11] Juan L. Gamella, Peter Bühlmann, and Jonas Peters. The causal chambers: Real physical systems as a testbed for AI methodology. *arXiv preprint arXiv:2404.11341*, 2024.

[12] Noam Gat and contributors. lm-format-enforcer, 2024. URL `https://github.com/noamgat/lm-format-enforcer`. A Python library for enforcing output formats in language models.

[13] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[14] Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.

[15] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, pages 16115–16152. PMLR, 2024.

[16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[17] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[19] Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1):388–427, 2021.

[20] Rob Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Australia, 3rd edition, 2021.

[21] Rob Hyndman, Anne B Koehler, J Keith Ord, and Ralph D Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media, 2008.

[22] 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. Time-LLM: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Unb5CVPtae`.

[23] Yaxuan Kong, Yiyuan Yang, Shiyu Wang, Chenghao Liu, Yuxuan Liang, Ming Jin, Stefan Zohren, Dan Pei, Yan Liu, and Qingsong Wen. Position: Empowering time series reasoning with multimodal llms. *arXiv preprint arXiv:2502.01477*, 2025.

[24] Kristina L Kupferschmidt, James Requiema, Mya Simpson, Zohrah Varsallay, Ethan Jackson, Cody Kupferschmidt, Sara El-Shawa, and Graham W Taylor. Food for thought: How can machine learning help better predict and understand changes in food prices? *arXiv preprint arXiv:2412.06472*, 2024.

[25] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6555–6565, 2024.

[26] Hunter Lightman, Vanya Kosaraju, Yura Burda, Harri Edwards, Benjamin Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[27] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4): 1748–1764, 2021.

[28] Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-MMD: A new multi-domain multimodal dataset for time series analysis. *arXiv preprint arXiv:2406.08627*, 2024.

[29] Haoxin Liu, Zhiyuan Zhao, Jindong Wang, Harshavardhan Kamarthi, and B. Aditya Prakash. LSTPrompt: Large language models as zero-shot time series forecasters by long-short-term prompting. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7832–7840, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.466. URL `https://aclanthology.org/2024.findings-acl.466/`.

[30] Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM on Web Conference 2024*, pages 4095–4106, 2024.

[31] David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6150–6160, Red Hook, NY, USA, 2018. Curran Associates Inc.

[32] Mike A Merrill, Mingtian Tan, Vinayak Gupta, Tom Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. *arXiv preprint arXiv:2404.11757*, 2024.

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

[34] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In *The Thirteenth International Conference on Learning Representations*, 2024.

[35] Martin Peterson. *An Introduction to Decision Theory*. Cambridge Introductions to Philosophy. Cambridge University Press, second edition, 2017. doi: 10.1017/9781316585061.

[36] Willa Potosnak, Cristian Challu, Mononito Goswami, Michał Wiliński, Nina Żukowska, and Artur Dubrawski. Implicit reasoning in deep time series forecasting. *arXiv preprint arXiv:2409.10840*, 2024.

[37] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Vincent Hassen, Anderson Schneider, et al. Lag-Llama: Towards foundation models for time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

[38] James Requeima, John Bronskill, Dami Choi, Richard E Turner, and David Duvenaud. LLM processes: Numerical predictive distributions conditioned on natural language. *arXiv preprint arXiv:2405.12856*, 2024.

[39] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2019.07.001. URL `https://www.sciencedirect.com/science/article/pii/S0169207019301888`.

[40] Manajit Sengupta, Yu Xie, Anthony Lopez, Aron Habte, Galen Maclaurin, and James Shelby. The national solar radiation data base (NSRDB). *Renewable and sustainable energy reviews*, 89: 51–60, 2018.

[41] 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. *ACM SIGKDD Explorations Newsletter*, 26(2):109–118, 2025.

[42] U.S. Bureau of Labor Statistics. Unemployment rate [various locations], 2024. URL `https://fred.stlouisfed.org/`. Accessed on 2024-08-30, retrieved from FRED.

[43] Ville de Montréal. Interventions des pompiers de montréal, 2020. URL `https://www.donneesquebec.ca/recherche/dataset/vmtl-interventions-service-securite-incendie-montreal`. Updated on 2024-09-12, accessed on 2024-09-13.

[44] Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. *arXiv preprint arXiv:2412.11376*, 2024.

[45] Chengsen Wang, Qi Qi, Zhongwen Rao, Lujia Pan, Jingyu Wang, and Jianxin Liao. Chronosteer: Bridging large language model and time series foundation model via synthetic data. *arXiv preprint arXiv:2505.10083*, 2025.

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[47] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: a survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6778–6786, 2023.

[48] Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, et al. Context is key: A benchmark for forecasting with essential textual information. In *ICML*, 2025.

[49] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

[50] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

[51] Zhijian Xu, Yuxuan Bian, Jianyuan Zhong, Xiangyu Wen, and Qiang Xu. Beyond trend and periodicity: Guiding time series forecasting with textual cues. *arXiv preprint arXiv:2405.13522*, 2024.

[52] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *CoRR*, 2024.

[53] Xiyuan Zhang, Boran Han, Haoyang Fang, Abdul Fatir Ansari, Shuai Zhang, Danielle C Maddix, Cuixiong Hu, Andrew Gordon Wilson, Michael W Mahoney, Hao Wang, et al. Does multimodality lead to better time series forecasting? *arXiv preprint arXiv:2506.21611*, 2025.

[54] Yunkai Zhang, Yawen Zhang, Ming Zheng, Kezhen Chen, Chongyang Gao, Ruian Ge, Siyuan Teng, Amine Jelloul, Jinmeng Rao, Xiaoyuan Guo, Chiang-Wei Fang, Zeyu Zheng, and Jie Yang. Insight miner: A large-scale multimodal model for insight mining from time series. In *NeurIPS 2023 AI for Science Workshop*, 2023. URL `https://openreview.net/forum?id=E1khscdUdH`.

[55] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[56] Xin Zhou, Weiqing Wang, Francisco J Baldán, Wray Buntine, and Christoph Bergmeir. Motime: A dataset suite for multimodal time series forecasting. *arXiv preprint arXiv:2505.15072*, 2025.

Appendix

Table of Contents

A Related Work

A.1 Large Models for Forecasting

Historically, classical methods such as ETS, ARIMA, and ensembles of such models have been at the cornerstone of time series forecasting [21, 4]. Following the explosion of deep learning in modalities such as vision and language, such methods were explored for forecasting tasks, starting with RNN-based and LSTM-based models [19, 39], followed by transformer methods [27, 50, 55, 8, 47, 33, 3]. Recently, following the success of pretrained large models in language [5], the time series community has also proposed foundation models for forecasting tasks [37, 15, 49, 2] pretrained on a large amount of time series data, and shown that they can output strong forecasts on unseen datasets zero-shot, outperforming models trained on those datasets. Research in large models for forecasting continues to grow, with efforts on better understanding their capabilities [36] and limitations [25]. A separate stream of research has explored using large language models (LLMs) for forecasting tasks. Gruver et al. [17] propose LLMTime, where an LLM is prompted to autoregressively generate a digit for each timestep in the prediction horizon of a forecasting task. They demonstrate surprising performance using LLMs for forecasting compared to time series models trained on specific datasets. Requeima et al. [38] improve this method to propose LLM Processes, where they show that the exact formatting of the prompt as well as the scaling factor used with the time series both matter. [29] propose LSTPrompt, which uses chain-of-thought methods to improve the performance of LLMs in quantitative forecasting tasks. As LLMs continue to improve regularly, several works have continued exploring the value of LLMs and limitations in forecasting [41]. Our work is related to these as we also aim to improve the forecasting performance of LLMs, however we operate in a setting where quantitative forecasting capabilities are insufficient, and tasks require models to understand the textual context to succeed in. Thereby, this setting of context-aided forecasting comes with different challenges and requires different capabilities from models [48], which we study with our methods.

A.2 Context-aided Forecasting Methods

One key capability that LLMs offer is the ability to condition on complementary side-information in text [22, 30, 44, 51, 28]. Jin et al. [22] propose Time-LLM, a multimodal model trained to use dataset-level metadata in addition to historical data for forecasting. The authors propose to use an LLM to encode the metadata and a transformer architecture to process time series, and train them jointly. Xu et al. [51], Liu et al. [28] propose similar multimodal architectures that can be trained for forecasting tasks on a specific dataset, where additional text information is available per time series window. Liu et al. [30] expand the setting to a multi-dataset setup and propose objective functions that allow better training on multiple time series based on their textual metadata, while preventing "domain confusion". As opposed to training a time series model from scratch, Wang et al. [45] propose an architecture that uses a pretrained time series foundation model in conjunction with an LLM in a similar setup, proposing to only train adapters between them. Wang et al. [44] adapt a purely-LLM-based approach to this, fine-tuning LLMs such as Qwen-7B on dataset-specific context-aided forecasting tasks, demonstrating the value of pure-LLM approaches. However, all these above methods involve training the model, which specializes the model to perform well solely on the time series and contexts that it is trained on [53]. Gruver et al. [17], Requeima et al. [38], Williams et al. [48], Merrill et al. [32] explore a different, zero-shot forecasting setup where the goal is to perform well on a diverse range of contexts and time-series. The focus in this setting is on how well models can use unambiguous, relevant context to succeed in forecasting scenarios, instead of on specializing models to specific scenarios [45]. Both Gruver et al. [17], Requeima et al. [38] demonstrate with preliminary results the ability of LLMs to successfully condition on textual information. Merrill et al. [32] evaluate a series of LLMs on context-aided forecasting tasks generated by GPT-4, and show that there is a huge gap between the performance of humans and that of LLMs on these tasks. To study the context-aided forecasting abilities of LLMs systematically, Williams et al. [48] propose a real-world evaluation benchmark of 71 zero-shot context-aided forecasting tasks across 8 different domains, each of which requires models to necessarily use the textual context to succeed in. The authors evaluate a range of LLMs zero-shot, with the LLMP method of Requeima et al. [38], and a faster and simpler prompting method they propose called Direct Prompt (DP), and demonstrate promising results with different LLMs. Our work builds on these results, going beyond early work on naïve direct prompting [48] and exploring variants which, as we demonstrate, can offer complementary advantages and reveal interesting insights into model capabilities, while significantly improving their performance.

B Additional Details on the Experimental Protocol

B.1 The RCRPS metric

We use the Region-of-Interest CRPS (RCRPS) metric to evaluate context-aided forecasting performance [48], which modifies the CRPS metric [13] to prioritize context-sensitive windows and accounts for constraint satisfaction. Given an inferred forecast distribution $\widetilde{\mathbf{X}}_F$ and a ground truth $\mathbf{x}_F$, the RCRPS metric is defined as:

$$\mathrm{RCRPS}(\widetilde{\mathbf{X}}_F,\mathbf{X}_F) = \alpha \cdot \left[\frac{1}{2|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathrm{CRPS}\big(\widetilde{X}_i, x_i\big) + \frac{1}{2|\neg\mathcal{I}|} \sum_{i \in \neg\mathcal{I}} \mathrm{CRPS}\big(\widetilde{X}_i, x_i\big) + \beta \cdot \mathrm{CRPS}\big(v_{\mathbf{C}}(\widetilde{\mathbf{X}}_F), 0\big) \right],$$

where the terms respectively account for the CRPS inside the RoI, the CRPS outside of the RoI, and the constraint violation penalty. The α term is a task-dependent normalization factor to make the RCRPS scale-independent, while β is a scaling factor that controls the impact of constraint violation on the score; we use $\beta = 10$ in our experiments as used in Williams et al. [48].

C Additional Results with the Direct Prompt (DP) Method

C.1 Aggregate Results of models

Results of various models with Direct Prompt (DP), with and without context are given in Table 3.

Model	Without Context	With Context
Qwen2.5-0.5B-Inst	**0.404** ± **0.028**	0.592 ± 0.027
Qwen2.5-1.5B-Inst	0.631 ± 0.039	**0.616** ± **0.018**
Qwen2.5-3B-Inst	0.513 ± 0.039	**0.424** ± **0.017**
Qwen2.5-7B-Inst	0.610 ± 0.011	**0.401** ± **0.006**
Qwen2.5-14B-Inst	0.551 ± 0.007	**0.247** ± **0.006**
Qwen2.5-32B-Inst	0.607 ± 0.008	**0.397** ± **0.008**
Qwen2.5-72B-Inst	0.549 ± 0.009	**0.202** ± **0.009**
Llama3.2-1B-Inst	0.481 ± 0.028	**0.396** ± **0.027**
Llama3.2-3B-Inst	0.950 ± 0.041	**0.687** ± **0.025**
Llama3-8B-Inst	0.758 ± 0.009	**0.543** ± **0.026**
Llama3.3-70B-Inst	0.700 ± 0.009	**0.230** ± **0.006**
Llama3.1-405B-Inst	0.686 ± 0.011	**0.173** ± **0.003**
Chronos-Large	0.492 ± 0.004	-
Lag-Llama	0.382 ± 0.011	-
Arima	0.636 ± 0.014	-

Table 3: Aggregate Results (RCRPS) of models on the CiK benchmark.

C.2 Results of models on various kinds of tasks

Results of various models with Direct Prompt (DP), with and without context, partitioned by different kinds of tasks are given in Table 4.

	ROI		non-ROI		Full ROI		Constraints	
Model	Without Context	With Context	Without Context	With Context	Without Context	With Context	Without Context	With Context
Llama3.2-1B-Inst	0.357 ± 0.018	**0.336** ± **0.026**	**0.236** ± **0.018**	0.248 ± 0.026	0.607 ± 0.045	**0.467** ± **0.041**	0.604 ± 0.064	**0.275** ± **0.092**
Llama3.2-3B-Inst	0.832 ± 0.118	**0.281** ± **0.013**	0.769 ± 0.030	**0.162** ± **0.013**	1.022 ± 0.048	**1.004** ± **0.040**	**0.613** ± **0.075**	1.030 ± 0.090
Llama3-8B-Inst	0.336 ± 0.017	**0.255** ± **0.008**	0.239 ± 0.017	**0.163** ± **0.008**	1.078 ± 0.009	**0.771** ± **0.043**	0.460 ± 0.199	**0.169** ± **0.172**
Qwen2.5-1.5B-Inst	0.327 ± 0.009	**0.317** ± **0.020**	**0.142** ± **0.009**	0.224 ± 0.020	0.900 ± 0.065	**0.851** ± **0.026**	**0.379** ± **0.242**	0.706 ± 0.147
Qwen2.5-7B-Inst	0.520 ± 0.008	**0.285** ± **0.006**	**0.157** ± **0.008**	0.164 ± 0.006	0.794 ± 0.018	**0.521** ± **0.009**	0.476 ± 0.041	**0.470** ± **0.078**
Qwen2.5-14B-Inst	0.376 ± 0.008	**0.162** ± **0.005**	0.155 ± 0.008	**0.146** ± **0.005**	0.745 ± 0.010	**0.310** ± **0.010**	0.473 ± 0.019	**0.039** ± **0.015**
Qwen2.5-32B-Inst	0.537 ± 0.003	**0.116** ± **0.001**	0.152 ± 0.003	**0.140** ± **0.001**	0.786 ± 0.014	**0.580** ± **0.013**	0.503 ± 0.031	**0.479** ± **0.019**
Llama3.3-70B-Inst	0.531 ± 0.010	**0.105** ± **0.003**	**0.147** ± **0.010**	0.182 ± 0.003	0.945 ± 0.014	**0.289** ± **0.011**	0.475 ± 0.031	**0.000** ± **0.024**
Llama3.1-405B-Inst	0.537 ± 0.002	**0.126** ± **0.004**	**0.147** ± **0.002**	0.150 ± 0.004	0.920 ± 0.019	**0.196** ± **0.005**	0.478 ± 0.038	**0.004** ± **0.009**
Qwen2.5-3B-Inst	0.280 ± 0.006	**0.269** ± **0.015**	**0.155** ± **0.006**	0.186 ± 0.015	0.713 ± 0.065	**0.558** ± **0.027**	**0.087** ± **0.147**	0.234 ± 0.056
Qwen2.5-72B-Inst	0.530 ± 0.001	**0.115** ± **0.004**	0.141 ± 0.001	**0.138** ± **0.004**	0.695 ± 0.015	**0.253** ± **0.015**	0.513 ± 0.034	**0.032** ± **0.028**
Qwen2.5-0.5B-Inst	**0.249** ± **0.005**	0.339 ± 0.010	0.149 ± 0.005	**0.129** ± **0.010**	**0.544** ± **0.046**	0.836 ± 0.046	0.557 ± 0.104	**0.243** ± **0.103**
Chronos-Large	0.536 ± 0.003		0.115 ± 0.003		0.605 ± 0.006		0.487 ± 0.010	
Lag-Llama	0.224 ± 0.005		0.202 ± 0.005		0.497 ± 0.018		0.204 ± 0.037	
Arima	0.272 ± 0.004		0.159 ± 0.004		0.921 ± 0.023		0.843 ± 0.050	

Table 4: Aggregate Results (RCRPS) of models on various groups of tasks from the CiK benchmark.

D Additional Details on ReDP

D.1 ReDP Prompt

We use the following prompt for the ReDP method, where **{history}** is replaced by the respective numerical history for the task instance in the format (timestamp, value), **{context}** is replaced by the respective textual context for the task instance, and ((**pred_time**)) is replaced with the prediction timesteps.

```
I have a time series forecasting task for you.

Here is some context about the task. Make sure to factor in any background
    knowledge,
satisfy any constraints, and respect any scenarios.
<context>
{context}
</context>

Here is a historical time series in (timestamp, value) format:
<history>
{history}
</history>

You are tasked with predicting the value at the following timestamps: {pred_time}.

First, within <reason> and </reason> tags, walk-through step-by-step how you
    would incorporate each piece of the context to improve your forecast. If you
    think any of the context is irrelevant, please indicate.

Next, return your forecast in (timestamp, value) format in between <forecast> and
    </forecast> tags.
Do not include any other information (e.g., comments) in the forecast.
```

One could use constrained decoding tools such as lm-format-enforcer [12] and XGrammar [7] to constrain the output format, however we found that using using constrained decoding with free-form text (between the <reasoning> and </reasoning>) was very slow, taking several hours for a single instance and at times not completing. Therefore, we do not use any constrained decoding and instead retry 15 times if a model fails to output in the specified format.

D.2 Reasoning Quality Analysis - Protocol Details

Tasks considered for the analysis. For the analysis, we consider the 20 context-aided forecasting tasks from the CiK benchmark [48] that have a region-of-interest (RoI) indicated. We consider only these tasks from the benchmark, due to the following reasons.

(i) **There exists only a single ground truth reasoning for these tasks.** For e.g. in the ElectricityIncreaseInPredictionTask, the only correct reasoning from the context would be to multiply the usual consumption for each hour affected by the heat wave, by the amount specified. In the other tasks where the full prediction window is the region of interest and the context does not specify a targeted region in the prediction window, several possible deductions can be made from the context that can help produce better forecasts [48]. Thereby, obtaining a single gold-standard reasoning trace that covers all these deductions is more difficult than in the former case. Evaluating a model's reasoning trace with such a gold-standard is also difficult, as the model's reasoning may partially be correct, which complicates evaluation.

(ii) **Measuring meaningful improvement with context is straight-forward in the case where there is a region of interest**. This is because originally, these tasks were created such that obtaining high accuracy in the region of interest is impossible without the context (as the data in these regions are modified appropriately, according to the context). Thereby, any improvement in this region of interest with context would mean that the model applied its reasoning correctly to obtain better performance: we conclude with an empirical analyses any improvement of 50% in this region as an improvement with context. On the contrary, in tasks where the entire prediction region is the region of interest which are more difficult tasks than the former [48], while models can improve their forecast with context, we found that the amount of percentage improvement required to conclude that the model has Due to these complications, we leave out these tasks from the analysis.

Obtaining gold-standard reasoning traces. For the reasoning quality analysis, we use the following prompt to obtain the gold-standard reasoning for the tasks considered in the analysis,

```
You are a forecasting expert. Given the following information:

CONTEXT:
{context}

Please provide a concise reasoning trace (one sentence) that explains how someone
    could logically produce a forecast based on the context.

Format your response as:
<reason>
[Your detailed reasoning here]
</reason>
```

We use GPT-4.1 to generate the gold-standard reasoning for all tasks. We manually verify the gold-standard reasoning traces and modify them if required.

Reasoning quality evaluation. Then, to compare a model's reasoning trace produced with ReDP Appendix D.1 (**{model_reasoning}** in the below prompt), with the gold-standard reasoning trace (**{ground_truth_reasoning}** in the below prompt), we use the following prompt, again with GPT4.1.

```
Compare these two reasoning traces for a forecasting task:

Model Reasoning:
{model_reasoning}

Ground Truth Reasoning:
{ground_truth_reasoning}

Question: Is the model reasoning aligned with the key points mentioned in ground
    truth reasoning approach?

Answer with exactly one word: YES or NO

<answer>YES/NO</answer>
```

We found that for the tasks considered in the analysis, the reasoning traces produced by a model were similar or many in cases exactly the same. Thereby, we only consider the reasoning trace corresponding to the first sample for the quality evaluation, for simplicity and to save costs (as the comparison needs to be done only once for a task instance). We will however release all the reasoning traces corresponding to all samples of a task, produced by a model. Extensions of the analysis methodology could look into comparing multiple reasoning traces using advancements in LLM-as-a-judge methodologies [10, 26, 18].

D.3 Performance of ReDP models

We plot in figures Figure 4 and Figure 5 the performance of ReDP models in the tasks considered for the analysis, compared to their respective performances using naive DP. We plot violin plots for each model, as different model fails in different tasks, precluding us from producing aggregate results on all tasks. Overall, we find that in many tasks, ReDP can slightly improve performance, but not significantly. In some tasks, ReDP even degrades performance. As the focus of the work is on evaluating the reasoning quality of the models through the ReDP method, we do not focus on the performance improvements or deteriorations that ReDP brings over DP.

D.4 Gold-standard reasoning traces

We release in Table 5 the gold-standard reasoning traces for the tasks considered for the analysis.

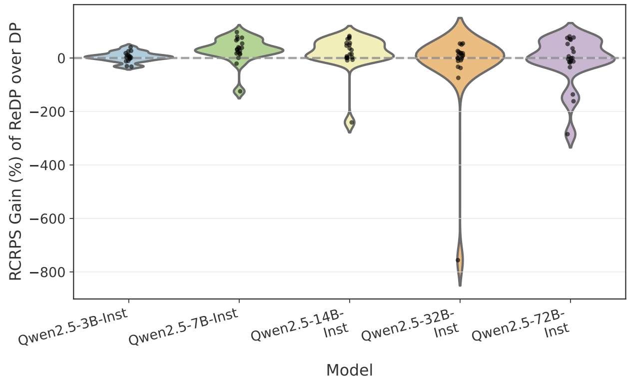


Figure 4: Distribution of improvement in RCRPS with ReDP over DP with Qwen models on the tasks considered for the analysis.

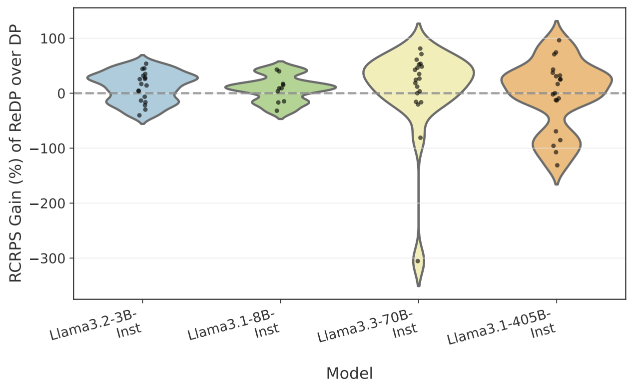


Figure 5: Distribution of improvement in RCRPS with ReDP over DP with Llama models on the tasks considered for the analysis.

Table 5: Gold Standard Reasoning for each task considered for the reasoning quality analysis.

Task Name	**Context**	**Ground Truth Reasoning**
ATMBuildingClosed Task	Background: This is the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England. Constraints: None. Scenario: Consider that the building which contains the ATM is closed from 1996-11-24 00:00:00, for 10 days.	Since the building housing the ATM is closed for 10 days starting 1996-11-24, a logical forecast would set the number of cash withdrawals to zero for that period, as no one can access the ATM during the closure.
ATMUnderPeriodic MaintenanceTaskWit hConclusion	Background: This is the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England. The ATM was under maintenance for 7 days, periodically every 15 days, starting from 1996-08-12 00:00:00, resulting in no withdrawals recorded. Assume that the ATM will not be in maintenance in the future. Constraints: None. Scenario: None.	A logical forecast can be produced by analyzing the historical withdrawal patterns while excluding data from the maintenance periods, then extrapolating the underlying trend and seasonality to predict future withdrawals, given that maintenance interruptions will no longer occur.

Task Name	Context	Ground Truth Reasoning
ATMUnderPeriodicMaintenanceTaskWithConclusionLessExplicit	Background: This is the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England. The ATM was under maintenance for various periods, resulting in no withdrawals recorded. Assume that the ATM will not be in maintenance in the future. Constraints: None. Scenario: None.	A logical forecast can be produced by identifying and removing periods of maintenance (with zero withdrawals) from the historical data to estimate the typical withdrawal rate during operational periods, then projecting this rate forward under the assumption that the ATM will remain operational.
ATMUnderPeriodicMaintenanceTaskWithoutConclusion	Background: This is the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England. The ATM was under maintenance for 7 days, periodically every 15 days, starting from 1996-08-12 00:00:00. Assume that the ATM will not be in maintenance in the future. Constraints: None. Scenario: None.	One could logically forecast future cash withdrawals by identifying and removing the recurring 7-day drops in activity caused by scheduled maintenance every 15 days, then modeling the underlying demand trend using the adjusted data to predict future withdrawals now that maintenance will no longer occur.
CashDepletedinATMScenarioTask	Background: This is the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England. Constraints: None. Scenario: Consider that cash is depleted in the ATM from 1996-11-24 00:00:00, for 10 days, resulting in no withdrawals during that period.	A forecaster could logically produce a forecast by identifying the 10-day period of zero withdrawals as an anomaly due to cash depletion, then modeling expected withdrawal counts for other periods based on historical data while treating the anomaly as missing or censored data rather than as indicative of typical demand.
DecreaseInTrafficInPredictionTask	Background: This is hourly traffic data. Constraints: None. Scenario: Suppose that there is an accident on the road and there is 20.0% of the usual traffic from 2024-01-18 06:00:00 for 5 hours.	To forecast traffic during the accident, multiply the usual hourly traffic by 0.2 for each hour from 2024-01-18 06:00:00 to 2024-01-18 11:00:00, since the accident reduces traffic to 20% of normal levels for 5 hours.
ElectricityIncreaseInPredictionTask	Background: This is the electricity consumption recorded in Kilowatt (kW) in city A. Constraints: None. Scenario: Suppose that there is a heat wave in city A from 2013-05-28 12:00:00 for 2 hours in city A, leading to excessive use of air conditioning, and 4 times the usual electricity being consumed.	To forecast electricity consumption during the heat wave, multiply the usual consumption for each hour between 2013-05-28 12:00:00 and 2013-05-28 14:00:00 by 4, since the scenario specifies consumption is quadrupled due to excessive air conditioning use.
ElectricityIncreaseInPredictionWithDistractorText	Background: This is the electricity consumption recorded in Kilowatt (kW) in city A. Constraints: None. Scenario: Suppose that there is a heat wave in city A from 2013-05-28 12:00:00 for 2 hours, leading to excessive use of air conditioning, and 4 times the usual electricity being consumed. A brief technical issue in the electricity grid caused a major dip of 75% in electricity consumption 2 weeks ago. This issue is not expected to happen again this week.	To forecast electricity consumption for the upcoming week, one should adjust the baseline usage by excluding the anomalous dip from two weeks ago, and account for a fourfold increase during the 2-hour heat wave period due to increased air conditioning demand.

Task Name	Context	Ground Truth Reasoning
ElectricityIncreaseInPredictionWithDistractorWithDates	Background: This is the electricity consumption recorded in Kilowatt (kW) in city A. Constraints: None. Scenario: There was a festival in neighbouring cities B and C that resulted in 10 times the usual electricity being consumed there from 2013-05-28 12:00:00 for 2 hours. But this did not affect electricity consumption in city A. Suppose that there is a heat wave in city A from 2013-05-28 12:00:00 for 2 hours, leading to excessive use of air conditioning, and 4 times the usual electricity being consumed.	To forecast electricity consumption in city A during the heat wave, one would identify the typical consumption for the relevant 2-hour period and multiply it by 4, since the scenario specifies that usage increases fourfold due to excessive air conditioning.
ElectricityIncreaseInPredictionWithSplitContext	Background: This is the electricity consumption recorded in Kilowatt (kW) in city A. Constraints: None. Scenario: Suppose that there is a heat wave in city A from 2013-05-28 12:00:00 for 2 hours, which would typically lead to excessive use of air conditioning, and 10 times the usual electricity being consumed. But in this case, residents sought to conserve energy and used lesser air conditioning, resulting in excessive usage of only 4 times the usual electricity.	A forecaster could estimate electricity consumption during the heat wave by identifying the usual consumption for the affected hours and multiplying it by 4, reflecting the adjusted behavior of residents who used less air conditioning than typical during such events.
ExplicitTrafficForecastTaskwithHolidaysInPredictionWindow	Background: This series contains the road occupancy rates on a freeway in the San Francisco Bay area. Note that 2024-07-04 is a holiday due to Independence Day. Note that traffic on this freeway typically reduces on holidays. Constraints: None. Scenario: None.	Given that July 4th is a holiday (Independence Day) and historical patterns show reduced freeway traffic on holidays, one could logically forecast a lower road occupancy rate for 2024-07-04 compared to typical weekdays by referencing past holiday data and general traffic trends.
ExplicitWithDatesAndDaysTrafficForecastTaskwithHolidaysInPredictionWindow	Background: This series contains the road occupancy rates on a freeway in the San Francisco Bay area. The days for which the forecast is required are Thursday 2024-07-04, Friday 2024-07-05, Saturday 2024-07-06. Note that 2024-07-04 is a holiday due to Independence Day. Note that traffic on this freeway typically reduces on holidays. Constraints: None. Scenario: None.	A logical forecast can be produced by analyzing historical occupancy rates for the same weekdays and dates, especially focusing on past Independence Days and adjacent days, to account for typical holiday traffic reductions and altered travel patterns.
ExplicitWithDaysTrafficForecastTaskwithHolidaysInPredictionWindow	Background: This series contains the road occupancy rates on a freeway in the San Francisco Bay area. The days for which the forecast is required are Thursday, Friday, Saturday. Note that 2024-07-04 is a holiday due to Independence Day. Note that traffic on this freeway typically reduces on holidays. Constraints: None. Scenario: None.	To forecast occupancy rates for Thursday, Friday, and Saturday, one could analyze historical occupancy patterns for the same days of the week, adjust for the expected reduction on Thursday due to the Independence Day holiday, and consider potential spillover effects on Friday and Saturday, as holiday travel and reduced commuter activity may influence these days as well.

Task Name	Context	Ground Truth Reasoning
ImplicitTrafficForecastTaskwithHolidaysInPredictionWindow	Background: This series contains the road occupancy rates on a freeway in the San Francisco Bay area. Note that traffic on this freeway typically reduces on holidays. Constraints: None. Scenario: None.	Given that road occupancy rates decrease on holidays, a logical forecast can be produced by identifying upcoming holidays in the calendar and adjusting expected occupancy rates downward for those dates relative to typical non-holiday levels.
IncreasedWithdrawal Scenario	Background: This is the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England. Constraints: None. Scenario: Suppose that there is a carnival from 1996-11-22 00:00:00, for 11 days leading to more people in the area, and 4 times the number of usual withdrawals during that period.	To forecast cash withdrawals during the carnival, multiply the usual daily withdrawal count by 4 for each of the 11 carnival days, while keeping the forecast for other days unchanged, as the scenario specifies a fourfold increase only during the event.
LongNewsElectricity IncreaseInPrediction Task	Background: This is the electricity consumption recorded in Kilowatt (kW) in city A. Constraints: None. Scenario: A sudden and intense heatwave struck the city, causing a dramatic surge in electricity consumption as residents sought refuge from the scorching temperatures. The extreme weather event, which began on 2013-05-28 12:00:00 and lasted for approximately 2 hours, saw temperatures soar to unprecedented levels. In response, citizens across the metropolitan area turned to their air conditioning units en masse, leading to a significant strain on the local power grid. According to the city's electricity provider, power consumption during the peak of the heatwave reached approximately 4 times the typical usage for this time of year. "We've never seen anything quite like this," said Jane Smith, spokesperson for PowerCity Utilities. "The sudden spike in demand pushed our systems to their limits." As the city recovers from this unprecedented power surge, experts are already discussing long-term solutions to manage similar situations in the future. These may include upgrades to the power grid, incentives for energy-efficient appliances, and the development of more robust emergency response protocols. For now, citizens are encouraged to stay hydrated, check on vulnerable neighbors, and use air conditioning responsibly as the community works together to beat the heat.	By analyzing the correlation between extreme heat events and surges in electricity consumption—specifically, noting that consumption during the heatwave was approximately four times the typical level for this period—one can forecast that future similar heatwaves are likely to cause comparable spikes in demand, unless mitigating measures are implemented.

Task Name	Context	Ground Truth Reasoning
MediumNewsElectricityIncreaseInPredictionTask	Background: This is the electricity consumption recorded in Kilowatt (kW) in city A. Constraints: None. Scenario: A sudden and intense heatwave struck the city, causing a dramatic surge in electricity consumption as residents sought refuge from the scorching temperatures. The extreme weather event, which began on 2013-05-28 12:00:00 and lasted for approximately 2 hours, saw temperatures soar to unprecedented levels. In response, citizens across the metropolitan area turned to their air conditioning units en masse, leading to a significant strain on the local power grid. According to the city's electricity provider, power consumption during the peak of the heatwave reached approximately 4 times the typical usage for this time of year. For now, citizens are encouraged to stay hydrated, check on vulnerable neighbors, and use air conditioning responsibly as the community works together to beat the heat.	By analyzing historical electricity consumption data for this time of year and multiplying the typical usage by four (as reported during the heatwave), one can estimate the likely electricity demand during similar future extreme heat events.
SensorMaintenanceInPredictionTask	Background: This series represents the occupancy rate (%) captured by a highway sensor. Constraints: None. Scenario: Consider that the meter will be offline for maintenance between 2024-01-18 08:00:00 and 2024-01-18 14:00:00, which results in zero readings.	The context mentions that the meter will be offline for maintenance between 2024-01-18 08:00:00 and 2024-01-18 14:00:00. During this period, one should forecast a value of 0%, as the meter will not be capturing any data.
SensorPeriodicMaintenanceTask	Background: This series represents the occupancy rate (%) captured by a highway sensor. The sensor was offline for maintenance every day between 08:00 and 14:00, which resulted in zero readings. Assume that the sensor will not be in maintenance in the future. Constraints: None. Scenario: None.	To produce a forecast, one could first remove the zero readings caused by maintenance periods to avoid bias, and then model the underlying occupancy rate pattern using the cleaned data, assuming future data will not be affected by maintenance downtime.
ShortNewsElectricityIncreaseInPredictionTask	Background: This is the electricity consumption recorded in Kilowatt (kW) in city A. Constraints: None. Scenario: A heatwave struck the city, which began on 2013-05-28 12:00:00 and lasted for approximately 2 hours, saw temperatures soar to unprecedented levels. According to the city's electricity provider, power consumption during the peak of the heatwave reached approximately 4 times the typical usage for this time of year.	By identifying the typical electricity consumption for late May at noon, multiplying this value by four to account for the reported heatwave surge, and applying this adjustment to the relevant 2-hour window, one can forecast the likely consumption during the heatwave event.

D.5 Examples of Correct and Incorrect Reasoning Traces

D.5.1 Task: *ElectricityIncreaseInPredictionTask*

Background: This series contains the amount of sunlight (in Watts per squared meter) arriving on a horizontal surface, for a location in Florida, United States. Over the previous 90 days, the maximum sunlight happened on average at 12:25:33.
Constraints: None
Scenario: None

Figure 6: Context

To forecast electricity consumption during the heat wave, multiply the usual consumption for each hour between 2013-05-28 12:00:00 and 2013-05-28 14:00:00 by 4, since the scenario specifies consumption is quadrupled due to excessive air conditioning use.

Figure 7: Gold Standard Reasoning Trace

To improve the forecast, we need to incorporate the given context. The scenario mentions a heat wave in city A from 2013-05-28 12:00:00 for 2 hours, leading to excessive use of air conditioning and 4 times the usual electricity being consumed.
First, we need to identify the usual electricity consumption pattern. From the historical data, we can observe the daily patterns and weekly patterns. However, we don't have any information about the seasonality or yearly patterns.
The heat wave is the most critical context for this forecasting task. We should consider that the electricity consumption will be 4 times the usual consumption during the heat wave (from 2013-05-28 12:00:00 to 2013-05-28 14:00:00).
For the timestamps before the heat wave (2013-05-28 04:00:00 to 2013-05-28 11:00:00), we can use the historical data to identify the daily pattern and forecast the values.
For the timestamps during the heat wave (2013-05-28 12:00:00 to 2013-05-28 14:00:00), we will multiply the forecasted values by 4.
For the timestamps after the heat wave (2013-05-28 15:00:00 onwards), we will assume that the electricity consumption returns to normal and use the historical data to forecast the values.
By incorporating the context, we can improve the accuracy of the forecast.

Figure 8: A Correct Reasoning Trace (Model: Llama3.3-70B-Inst)

First, I would incorporate the historical time series data into my model. I would split the data into training and testing sets, with the training set consisting of data from 2013-05-22 04:00:00 to 2013-05-27 03:00:00, and the testing set consisting of data from 2013-05-27 04:00:00 to 2013-05-27 23:00:00 and 2013-05-28 12:00:00 (since it's the start of the heatwave).
Next, I would incorporate the scenario that there is a heat wave in city A from 2013-05-28 12:00:00 for 2 hours, leading to excessive use of air conditioning, and 4 times the usual electricity being consumed. To incorporate this, I would analyze the historical data and identify patterns in electricity consumption that are likely to be affected by heatwaves, such as increased consumption during the peak sun hours (usually between 11am and 3pm). I would then adjust the model to account for these increased consumption patterns during the heatwave.
I would also consider the relevance of each piece of the context. In this case, I consider the historical data, the heatwave scenario, and the timestamps for which I need to make predictions as relevant.
Here's how I would implement the model:
I would use a time series forecasting model that is capable of handling seasonal and trend components, such as an ARIMA or a seasonal decomposition model. I would then use the historical data to train the model and make predictions for the future timestamps.
However, since the heatwave scenario changes the behavior of the system, I would need to adjust the predictions accordingly. I would do this by identifying the periods affected by the heatwave and adjusting the predicted values to reflect the increased consumption due to the heatwave.
Finally, I would evaluate the performance of the model on the testing data and refine the model as needed.
In terms of which of the context is irrelevant, I would say that the specific background information about city A is not directly relevant to the task at hand, although it could be useful for additional analysis.

Figure 9: A Correct Reasoning Trace (Model: Llama3.1-8B-Inst)

Given the provided historical data and the context of a heat wave causing a significant increase in electricity consumption, I will make adjustments to the forecast based on the sudden surge in demand during the heat wave period. Here's the step-by-step reasoning:
1. **Identify the Heat Wave Period**: From the history, identify the period around 2013-05-28 12:00:00 when electricity consumption spiked to 555.84 kW. This likely represents the start of the heat wave impact.
2. **Analyze the Trend Before and After the Heat Wave**: Analyze the trend before and after the identified heat wave period to understand the normal variation and any potential anomalies or patterns during the heat wave.
- **Before Heat Wave** (e.g., 2013-05-22 to 2013-05-27): Observe that values tend to stabilize around 400-600 kW. - **After Heat Wave** (e.g., 2013-05-28 onwards): Values increase significantly due to higher air conditioning usage, potentially reaching up to 700 kW.
3. **Forecast Adjustments**: - For the immediate post-heat wave period (2013-05-28), expect a continuation of the high values seen during the heat wave but with some normalization as it cools down. - After the initial spike, the forecast should follow the pre-heat wave trend. - During the subsequent days (2013-05-29 and beyond), the electricity consumption should return to a more stable level, possibly within the normal range observed before the heat wave. Based on this analysis, the adjusted forecasts can be made accordingly.

Figure 10: An Incorrect Reasoning Trace (Model: Qwen2.5-3B-Inst)

To improve the forecast, I will incorporate the context as follows:
1. **Background:** The data represents electricity consumption in Kilowatt (kW) in city A. This information is useful to understand the general trend and pattern of electricity consumption.
2. **Scenario:** There is a heat wave in city A from 2013-05-28 12:00:00 for 2 hours, leading to excessive use of air conditioning, and 4 times the usual electricity being consumed. This information is crucial to adjust the forecast for the specified time period. I will make sure to increase the forecasted values for 2013-05-28 12:00:00 and 2013-05-28 13:00:00 by a factor of 4.
To generate the forecast, I will use a simple moving average approach, considering the patterns and trends observed in the historical data.

Figure 11: An Incorrect Reasoning Trace (Model: Llama3.1-405B-Inst)

D.5.2 Task: ***SensorMaintenanceInPredictionTask***

Background: This series represents the occupancy rate (%) captured by a highway sensor.
Constraints: None.
Scenario: Consider that the meter will be offline for maintenance between 2024-01-18 08:00:00 and 2024-01-18 14:00:00, which results in zero readings.

Figure 12: Context

The context mentions that the meter will be offline for maintenance between 2024-01-18 08:00:00 and 2024-01-18 14:00:00. During this period, one should forecast a value of 0

Figure 13: Gold Standard Reasoning Trace

The provided historical data shows a daily pattern where occupancy rates increase significantly during morning hours, peak around midday, and then decrease gradually throughout the rest of the day. The meter will be offline for maintenance from 2024-01-18 08:00:00 to 2024-01-18 14:00:00, resulting in zero readings during this period. To account for this, we should predict normal values until 08:00 and set all predictions to zero during the maintenance window. After 14:00, we should resume forecasting based on the historical patterns.

Figure 14: A Correct Reasoning Trace (Model: Qwen2.5-14B-Inst)

To incorporate the context provided into the forecast:
1. **Background**: The series represents an occupancy rate captured by a highway sensor. This indicates we should expect patterns related to traffic flows, which are likely influenced by daily routines such as rush hours, commuting, and other regular activities.
2. **Meter Offline Scenario**: The meter will be offline for maintenance between 2024-01-18 08:00:00 and 2024-01-18 14:00:00, resulting in zero readings during this period. For the timestamps within this range, we must predict zero occupancy rates directly based on this constraint.
Given the data history, it shows significant variability, particularly around early morning and late evening times, suggesting peak and off-peak traffic periods. We can assume similar patterns will continue unless there's a specific reason to believe otherwise.
For the timestamps before and after the maintenance period, we can infer trends from the previous days. Observing the last few days (January 15 to January 17), we see that the occupancy rate typically peaks in the morning around 7 AM, drops significantly midday, and then has another small peak in the late afternoon/early evening. We will use these patterns to guide our predictions.

Figure 15: A Correct Reasoning Trace (Model: Qwen2.5-32B-Inst)

Given the provided historical data, the first step is to preprocess the data to handle the scenario where the meter is offline from 2024-01-18 08:00:00 to 2024-01-18 14:00:00. Since the readings are zero during this period, we need to identify patterns or trends before and after this period to forecast values accurately.
Step-by-Step Reasoning:
1. **Preprocess Data**: Identify the missing values due to the meter being offline. - From the history, we see that the readings from 2024-01-18 08:00:00 to 2024-01-18 14:00:00 are all zeros.
2. **Identify Patterns Before and After**: Analyze the trend before and after the offline period to make informed predictions. - Examine the trends from the day before (2024-01-17) and the next day (2024-01-19) to understand any seasonality or cyclical patterns.
3. **Forecast Values**: Use the identified patterns and possibly apply smoothing techniques like moving averages or seasonal decomposition to predict the missing values.
4. **Return Forecasts**: Return the predicted values for the requested timestamps.

Figure 16: An Incorrect Reasoning Trace (Model: Qwen2.5-3B-Inst)

1. **Understanding the Context**: The provided time series data represents an occupancy rate captured by a highway sensor. The sensor is expected to be offline from 2024-01-18 08:00:00 to 2024-01-18 14:00:00, which means the data for these timestamps will be missing. This gap needs to be accounted for in the forecasting model.
2. **Analyzing the Historical Data**: The historical data shows a range of occupancy rates, indicating a daily cycle. The data starts at a lower occupancy rate in the early hours and gradually increases to a peak in the late morning and early afternoon, then decreases as the day progresses.
3. **Model Selection**: Given the daily cycle pattern, an ARIMA model could be suitable. However, considering the specific missing data, a more comprehensive model like a Seasonal ARIMA (SARIMA) might be necessary to account for both the daily seasonality and the missing data.
4. **Handling Missing Data**: For the period during the maintenance (2024-01-18 08:00:00 to 2024-01-18 14:00:00), we will impute the values with a method that considers the surrounding data points to make the forecast as accurate as possible. We can use the average or median of the surrounding data points to fill in the gaps.
5. **Forecasting**: We will extend the model with the historical data and imputed values for the maintenance period. The forecast will then be continued beyond the maintenance period based on the identified pattern.

Figure 17: An Incorrect Reasoning Trace (Model: Qwen2.5-7B-Inst)

D.5.3 Task: ***CashDepletedinATMScenarioTask***

Background: This is the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England.
Constraints: None.
Scenario: Consider that cash is depleted in the ATM from 1996-11-24 00:00:00, for 10 days, resulting in no withdrawals during that period.

Figure 18: Context

A forecaster could logically produce a forecast by identifying the 10-day period of zero withdrawals as an anomaly due to cash depletion, then modeling expected withdrawal counts for other periods based on historical data while treating the anomaly as missing or censored data rather than as indicative of typical demand.

Figure 19: Gold Standard Reasoning Trace

To address this forecasting task, I will first assess the historical data pattern and then incorporate the provided context, which includes a scenario where the ATM was out of cash for ten days starting from 1996-11-24 00:00:00. The historical data shows variability, indicating a mix of daily and potentially weekly patterns. The scenario of the ATM being out of cash for ten days implies that for these days, there would be no withdrawals recorded. Following this event, it is reasonable to expect an increase in withdrawal numbers as the machine is replenished and users resume their normal withdrawal behaviors.
Given the lack of specific seasonal or trend data beyond day-to-day fluctuations, the forecast will be based on averaging recent withdrawal counts while considering the impact of the ATM's refill post-cash depletion. Since the exact replenishment amount and the timing of users resuming withdrawals are unknown, I'll use a conservative approach by estimating withdrawals to return to a typical level observed in the past few weeks, taking into account the likely higher activity after the refill.
For the period when the ATM was empty (1996-11-24 to 1996-12-03), I will predict zero withdrawals. After this period, withdrawals will be predicted based on a moving average of the recent non-zero values, adjusted for the expected increase due to replenishment.

Figure 20: A Correct Reasoning Trace (Model: Qwen2.5-32B-Inst)

Given the provided historical data, it appears there was a significant disruption in the time series on November 13, 1996, as the withdrawal count dropped significantly and remained low for the next several days. It is likely that this event had a strong influence on the subsequent values. To incorporate this information effectively:
1. **Identify the Abrupt Change**: The value drops sharply from 1996-11-13 to 1996-11-14 (from approximately 21.97 to 0.23). This drop suggests a sudden decrease in transactions which could be influenced by various factors such as an event or system downtime.
2. **Examine Patterns Before and After**: The immediate aftermath shows a sustained low level of transactions followed by a gradual increase. The drop and recovery might imply that once the disruption is resolved, activity will likely return to its previous levels but may not reach pre-disruption levels immediately due to residual effects.
3. **Forecasting Strategy**: - For the days directly affected (11/13-11/14), we can forecast based on the abrupt drop. Assuming the trend resumes after the initial shock, we can predict higher values compared to the pre-shock trend. - For the subsequent days, the forecast should reflect a gradual increase towards normalcy.

Figure 21: An Incorrect Reasoning Trace (Model: Qwen2.5-3B-Inst)

D.6 End-to-end Analyses Examples

D.6.1 Task: *ATMBuildingClosedTask*

Background: This is the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England.
Constraints: None.
Scenario: Consider that the building which contains the ATM is closed from 1996-11-24 00:00:00, for 10 days.

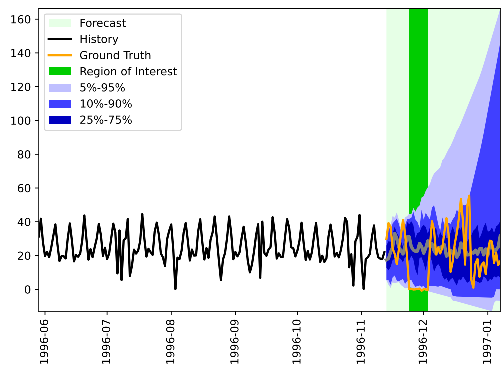
Figure 22: Context

Since the building housing the ATM is closed for 10 days starting 1996-11-24, a logical forecast would set the number of cash withdrawals to zero for that period, as no one can access the ATM during the closure.

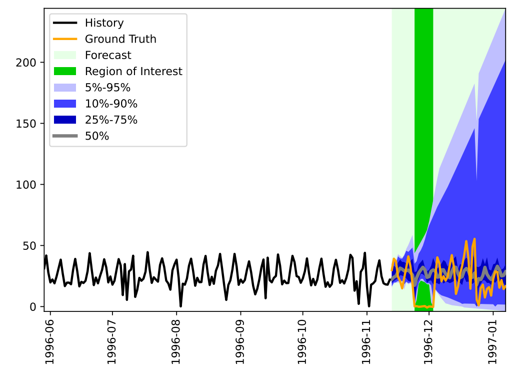
Figure 23: Gold Standard Reasoning Trace

First, we need to understand the historical data and identify any patterns or seasonal trends in the time series. In this case, we see that the number of cash withdrawals from the ATM is quite seasonal, with the highest values in the summer months and lowest values in the winter months.
Next, we need to incorporate the background information about the building being closed from 1996-11-24 00:00:00 for 10 days into our model. This could potentially impact the number of cash withdrawals during that period, so we need to account for this in our forecasting model.
We also need to consider the given historical data and ensure that our model is not overfitting to the training data. We can do this by using a technique such as cross-validation to evaluate the performance of our model on unseen data.
Another important consideration is the fact that the data is from England and we should consider the local seasonal patterns, if any.
Finally, we need to decide on a suitable forecasting model that can handle seasonal and non-seasonal trends, as well as the potential impact of the building closure on the data.

(a) Reasoning trace



(b) Forecast without context (RCRPS=0.434)

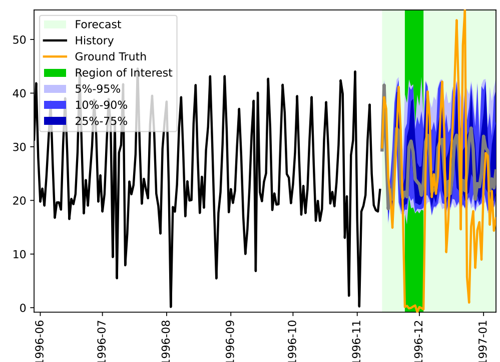


(c) Context-aided Forecast (RCRPS=0.565)

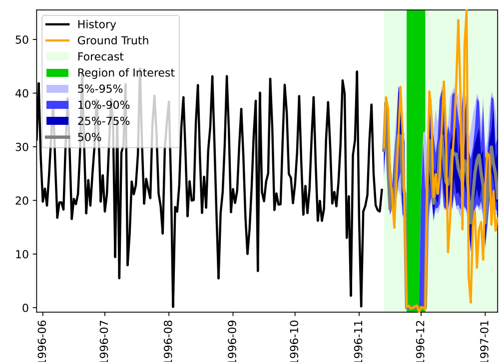
Figure 24: Model: Llama3.2-3B-Inst. The reasoning trace is evaluated as **Incorrect Reasoning**. The forecast changes by -30.1% and is hence evaluated as **No Improvement with Context**.

To improve the forecast, we need to consider the context provided. The building which contains the ATM is closed from 1996-11-24 00:00:00 for 10 days. This means that there will be no cash withdrawals during this period.
First, we need to identify the start and end dates of the closure. The start date is 1996-11-24 00:00:00, and the end date is 1996-12-04 00:00:00 (10 days after the start date).
Next, we need to adjust our forecast to reflect the closure. For the dates between 1996-11-24 00:00:00 and 1996-12-04 00:00:00, we should predict a value of 0, since there will be no cash withdrawals during this period.
For the dates after 1996-12-04 00:00:00, we can use a time series forecasting model, such as ARIMA or Prophet, to predict the values. We can train the model using the historical data provided.
The historical data shows a daily frequency, with some fluctuations in the values. We can use this data to identify patterns and trends, and then use these patterns to make predictions for the future dates.
Overall, the context provided is relevant to the forecast, as it gives us information about a specific event (the closure of the building) that will affect the time series.

(a) Reasoning trace



(b) Forecast without context (RCRPS=0.608)

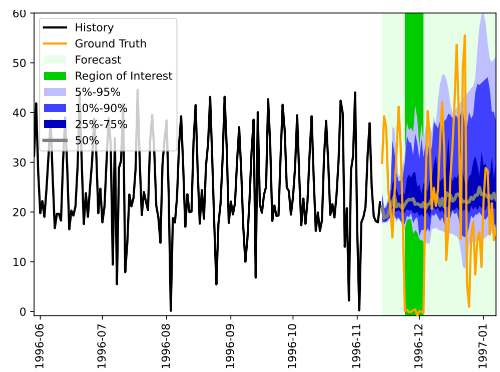


(c) Context-aided Forecast (RCRPS=0.013)

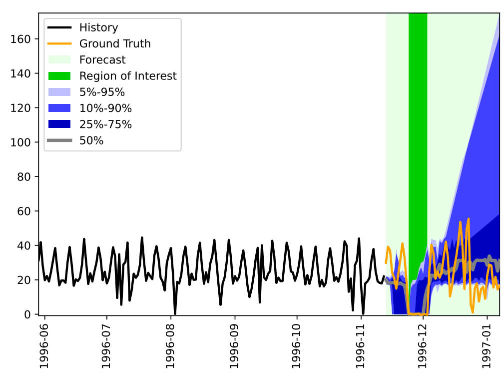
Figure 25: Model: Llama-3.3-70B-Inst. The reasoning trace is evaluated as **Correct Reasoning**. The forecast changes by 97.9% and is hence evaluated as **Improvement with Context**.

The context provides us with a scenario where the building containing the ATM will be closed from 1996-11-24 00:00:00 for 10 days. During this period, we can reasonably expect the ATM to be inaccessible and therefore no cash withdrawals will occur. This means that for the timestamps '1996-11-24 00:00:00' to '1996-11-27 00:00:00', we should predict zero values as the ATM will not be operational. For the remaining timestamps, we can use the historical data to forecast the expected withdrawal counts, taking into account the closure period. However, after the closure ends on 1996-11-28 00:00:00, the number of withdrawals should return to a level similar to the pre-closure trend, but it might take some time for usage to fully recover to normal levels.

(a) Reasoning trace



(b) Forecast without context (RCRPS=0.545)

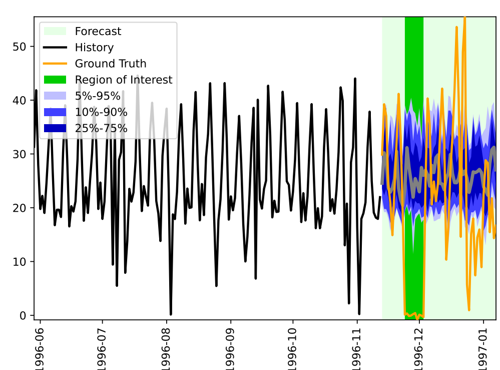


(c) Context-aided Forecast (RCRPS=0.060)

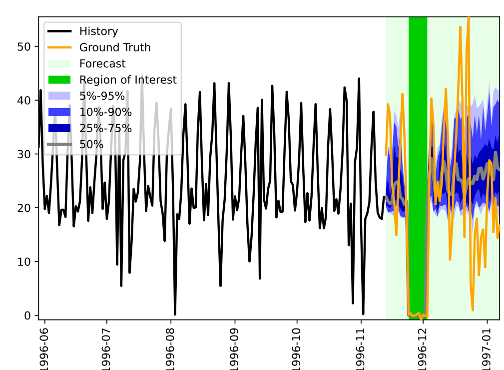
Figure 26: Model: Qwen2.5-14B-Inst. The reasoning trace is evaluated as **Correct Reasoning**. The forecast changes by 88.9% and is hence evaluated as **Improvement with Context**.

To improve my forecast, I will incorporate each piece of the context as follows:
1. **Background**: The data represents the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England. This context is relevant as it helps me understand the type of data I'm dealing with and potential seasonal or periodic patterns.
2. **Scenario**: The building containing the ATM is closed from 1996-11-24 00:00:00 for 10 days. This context is crucial as it directly affects the data during the forecast period. I will ensure that my forecast reflects the closure by predicting zero or significantly reduced cash withdrawals during this period.
Next, I will analyze the historical data to identify any patterns, trends, or seasonality that can inform my forecast.

(a) Reasoning trace



(b) Forecast without context (RCRPS=0.609)



(c) Context-aided Forecast (RCRPS=0.025)

Figure 27: Model: Llama3.1-405B-Inst. The reasoning trace is evaluated as **Correct Reasoning**. The forecast changes by 95.9% and is hence evaluated as **Improvement with Context**.

D.6.2 Task: *CashDepletedinATMScenarioTask*

Background: This is the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England.
Constraints: None.
Scenario: Consider that cash is depleted in the ATM from 1996-11-24 00:00:00, for 10 days, resulting in no withdrawals during that period.

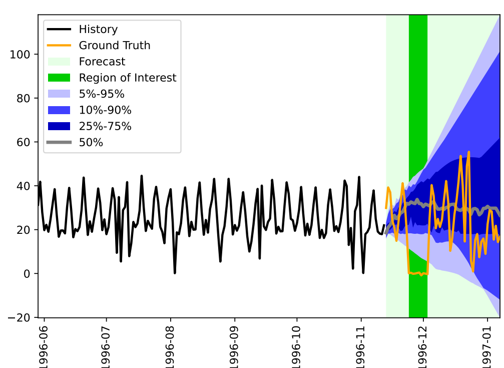
Figure 28: Context

A forecaster could logically produce a forecast by identifying the 10-day period of zero withdrawals as an anomaly due to cash depletion, then modeling expected withdrawal counts for other periods based on historical data while treating the anomaly as missing or censored data rather than as indicative of typical demand.

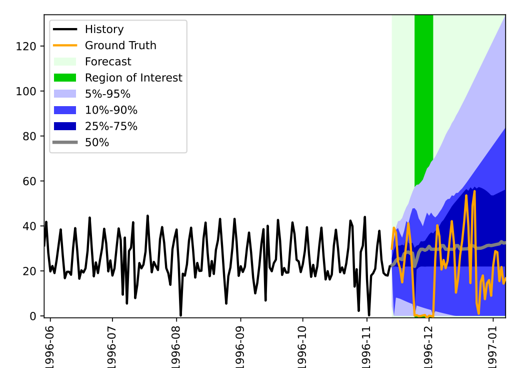
Figure 29: Gold Standard Reasoning Trace

Given the provided historical data, it appears there was a significant disruption in the time series on November 13, 1996, as the withdrawal count dropped significantly and remained low for the next several days. It is likely that this event had a strong influence on the subsequent values. To incorporate this information effectively:
1. **Identify the Abrupt Change**: The value drops sharply from 1996-11-13 to 1996-11-14 (from approximately 21.97 to 0.23). This drop suggests a sudden decrease in transactions which could be influenced by various factors such as an event or system downtime.
2. **Examine Patterns Before and After**: The immediate aftermath shows a sustained low level of transactions followed by a gradual increase. The drop and recovery might imply that once the disruption is resolved, activity will likely return to its previous levels but may not reach pre-disruption levels immediately due to residual effects.
3. **Forecasting Strategy**: - For the days directly affected (11/13-11/14), we can forecast based on the abrupt drop. Assuming the trend resumes after the initial shock, we can predict higher values compared to the pre-shock trend. - For the subsequent days, the forecast should reflect a gradual increase towards normalcy.

(a) Reasoning trace



(b) Forecast without context (RCRPS=0.653)

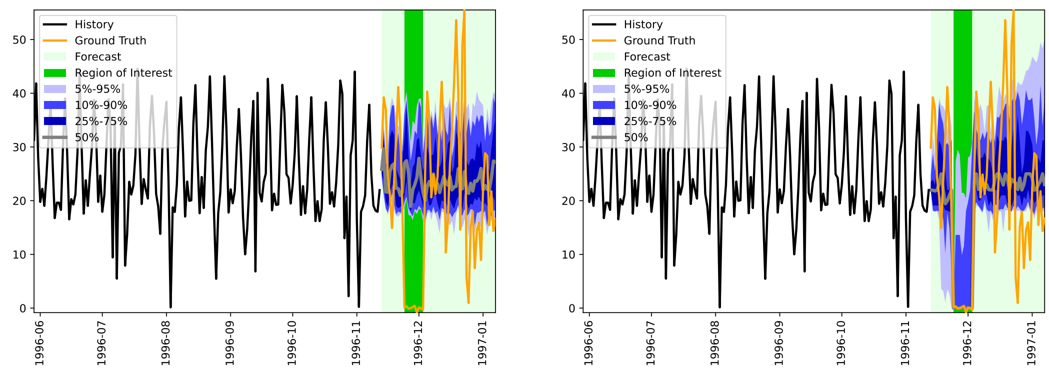


(c) Context-aided Forecast (RCRPS=0.493)

Figure 30: Model: Qwen2.5-3B-Inst. The reasoning trace is evaluated as **Incorrect Reasoning**. The forecast changes by 24.6% and is hence evaluated as **No Improvement with Context**.

To address this forecasting task, I will first assess the historical data pattern and then incorporate the provided context, which includes a scenario where the ATM was out of cash for ten days starting from 1996-11-24 00:00:00. The historical data shows variability, indicating a mix of daily and potentially weekly patterns. The scenario of the ATM being out of cash for ten days implies that for these days, there would be no withdrawals recorded. Following this event, it is reasonable to expect an increase in withdrawal numbers as the machine is replenished and users resume their normal withdrawal behaviors.
Given the lack of specific seasonal or trend data beyond day-to-day fluctuations, the forecast will be based on averaging recent withdrawal counts while considering the impact of the ATM's refill post-cash depletion. Since the exact replenishment amount and the timing of users resuming withdrawals are unknown, I'll use a conservative approach by estimating withdrawals to return to a typical level observed in the past few weeks, taking into account the likely higher activity after the refill.
For the period when the ATM was empty (1996-11-24 to 1996-12-03), I will predict zero withdrawals. After this period, withdrawals will be predicted based on a moving average of the recent non-zero values, adjusted for the expected increase due to replenishment.

(a) Reasoning trace



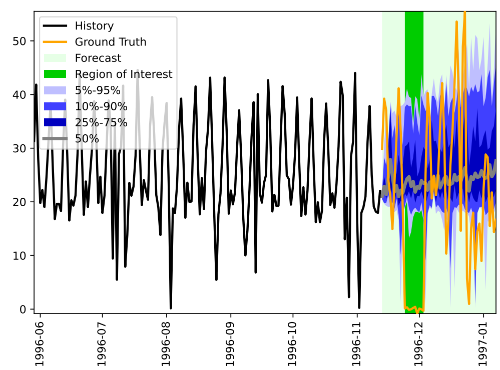
(b) Forecast without context (RCRPS=0.586)

(c) Context-aided Forecast (RCRPS=0.023)

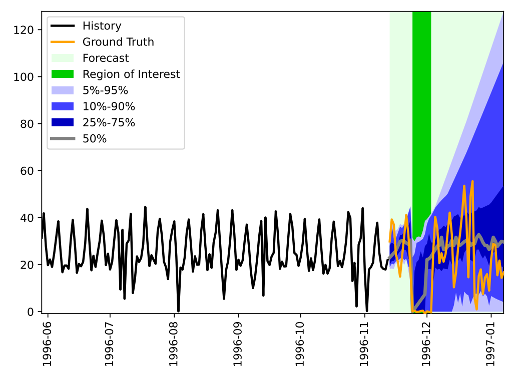
Figure 31: Model: Qwen2.5-32B-Inst. The reasoning trace is evaluated as **Correct Reasoning**. The forecast changes by 96.0% and is hence evaluated as **Improvement with Context**.

1. **Context Inclusion**: We need to consider the context of the task, which includes the historical data showing a significant drop in withdrawals from 1996-11-24 to 1996-11-33 due to ATM cash depletion. This drop needs to be factored into our forecasting model.
2. **Model Choice**: Given the need for robust time-series forecasting and the presence of a data gap, we will use a combination of exponential smoothing and ARIMA (AutoRegressive Integrated Moving Average) models. These methods are well-suited for handling missing data and incorporate both the seasonal and trend components of the time series.
3. **Handling the Gap**: For the period from 1996-11-24 to 1996-11-33, we will assume that the withdrawals were zero or, if necessary, we will backfill the missing values using a conservative estimate or interpolation techniques. This is critical for avoiding bias in our forecasting model.
4. **Seasonality and Trend Analysis**: We will analyze the time series for seasonality and trends. The historical data shows variability, with some periodic patterns, which we will capture in our model.
5. **Model Training**: We will train our model on the entire historical dataset, ensuring to handle the gap appropriately. This will help the model learn from the historic consistency and variability.
6. **Forecasting**: Finally, we will use the trained model to predict future values, ensuring to respect the natural patterns and handle the gap by predicting zeros for the known period.

(a) Reasoning trace



(b) Forecast without context (RCRPS=0.572)



(c) Context-aided Forecast (RCRPS=0.193)

Figure 32: Model: Qwen-2.5-7B-Inst. The reasoning trace is evaluated as **Correct Reasoning**. The forecast changes by 66.2% and is hence evaluated as **Improvement with Context**.

D.6.3 Task: *ElectricityIncreaseInPredictionTask*

Background: This series contains the amount of sunlight (in Watts per squared meter) arriving on a horizontal surface, for a location in Florida, United States. Over the previous 90 days, the maximum sunlight happened on average at 12:25:33.
Constraints: None
Scenario: None

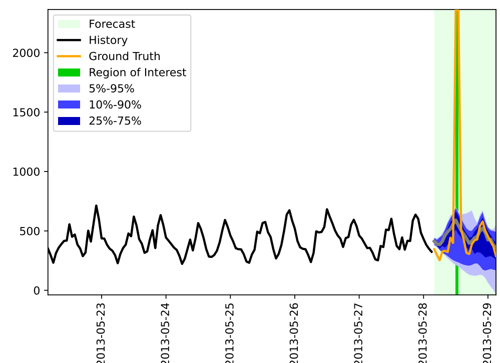
Figure 33: Context

To forecast electricity consumption during the heat wave, multiply the usual consumption for each hour between 2013-05-28 12:00:00 and 2013-05-28 14:00:00 by 4, since the scenario specifies consumption is quadrupled due to excessive air conditioning use.

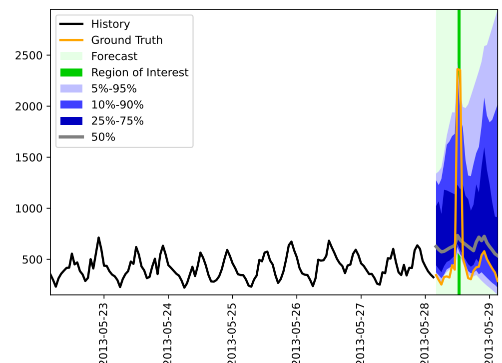
Figure 34: Gold Standard Reasoning Trace

To incorporate the given context into the time series forecasting task, we can follow these steps:
1. Preprocess the historical time series data: Clean and preprocess the historical data by converting all timestamps to a uniform format, handling missing values, and normalizing the data.
2. Identify the trend in the historical data: Use techniques such as linear regression, polynomial regression, or ARIMA to identify the trend in the historical data.
3. Account for seasonal patterns: Given that the data is hourly, it is likely that there are seasonal patterns in the data. We can use techniques such as seasonal decomposition or SARIMA models to account for these patterns.
4. Consider anomalies and outliers: Given that there was a heat wave in city A from 2013-05-28 12:00:00 for 2 hours, we can consider this as an anomaly in the data. We can use techniques such as ARIMA or Exponential Smoothing models to account for this anomaly.
5. Incorporate the heat wave anomaly into the model: We can use the heat wave anomaly as a basis to predict the consumption for the next two hours after the heat wave. This will require a model that can capture the anomalies and make predictions based on them.
6. Use the model to forecast future values: Once we have trained the model with the historical data and the heat wave anomaly, we can use it to forecast future values for the next 24 hours.

(a) Reasoning trace



(b) Forecast without context (RCRPS=0.068)

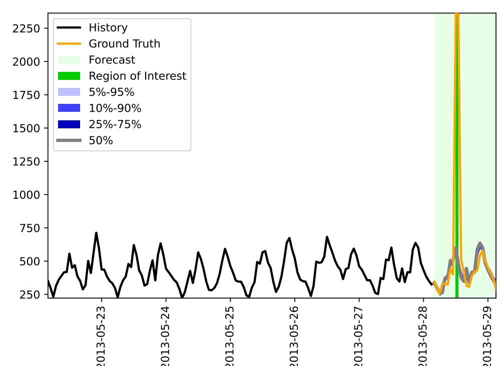


(c) Context-aided Forecast (RCRPS=0.038)

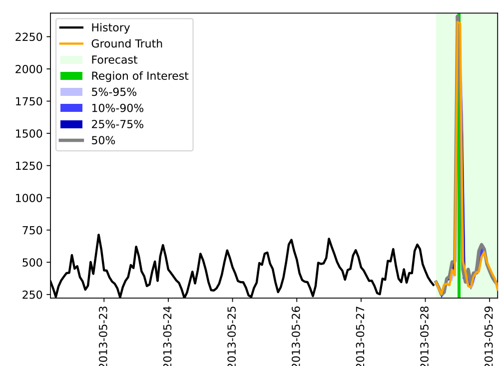
Figure 35: Model: Llama3.2-3B-Inst. The reasoning trace is evaluated as **Incorrect Reasoning**. The forecast changes by 43.9% and is hence evaluated as **No Improvement with Context**.

1. **Understand the Historical Data**: Analyze the historical data to identify patterns, trends, and seasonality. The data shows hourly electricity consumption over a period, which can be used to understand typical consumption patterns.
2. **Identify the Heat Wave Period**: The heat wave is specified to occur from 2013-05-28 12:00:00 for 2 hours. During this period, the electricity consumption is expected to be 4 times the usual amount. This is a critical piece of information that needs to be incorporated into the forecast.
3. **Adjust for the Heat Wave**: For the timestamps 2013-05-28 12:00:00 and 2013-05-28 13:00:00, the forecasted values should be adjusted to 4 times the expected consumption based on historical data. For the other timestamps, the forecast can be made based on historical patterns, considering the usual daily and weekly trends.
4. **Use Historical Data for Forecasting**: Use the historical data to predict the values for the remaining timestamps. For example, if the consumption on a similar day (e.g., 5 days prior) was around 400 kW at 04:00:00, the forecast for 2013-05-28 04:00:00 can be based on this value.
5. **Incorporate Seasonality**: Ensure that the forecast takes into account the typical increase in consumption during the afternoon and evening hours, which is evident from the historical data.
6. **Relevance of Context**: The context about the heat wave is highly relevant and must be factored into the forecast. The historical data is also crucial for making accurate predictions.

(a) Reasoning trace



(b) Forecast without context (RCRPS=0.072)

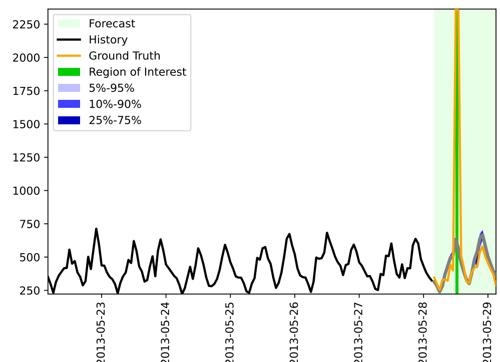


(c) Context-aided Forecast (RCRPS=0.008)

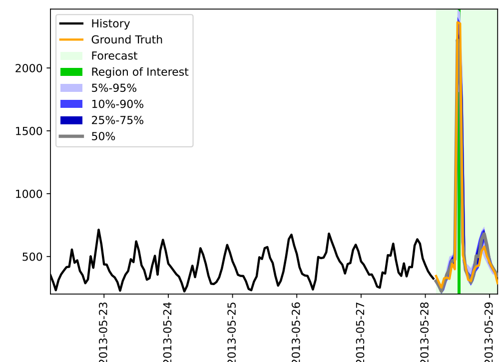
Figure 36: Model: Qwen-2.5-72B-Inst. The reasoning trace is evaluated as **Correct Reasoning**. The forecast changes by 89.3% and is hence evaluated as **Improvement with Context**.

To improve the forecast, we need to incorporate the given context. The scenario mentions a heat wave in city A from 2013-05-28 12:00:00 for 2 hours, leading to excessive use of air conditioning and 4 times the usual electricity being consumed.
First, we need to identify the usual electricity consumption pattern. From the historical data, we can observe the daily patterns and weekly patterns. However, we don't have any information about the seasonality or yearly patterns.
The heat wave is the most critical context for this forecasting task. We should consider that the electricity consumption will be 4 times the usual consumption during the heat wave (from 2013-05-28 12:00:00 to 2013-05-28 14:00:00).
For the timestamps before the heat wave (2013-05-28 04:00:00 to 2013-05-28 11:00:00), we can use the historical data to identify the daily pattern and forecast the values.
For the timestamps during the heat wave (2013-05-28 12:00:00 to 2013-05-28 14:00:00), we will multiply the forecasted values by 4.
For the timestamps after the heat wave (2013-05-28 15:00:00 onwards), we will assume that the electricity consumption returns to normal and use the historical data to forecast the values.
By incorporating the context, we can improve the accuracy of the forecast.

(a) Reasoning trace



(b) Forecast without context (RCRPS=0.069)



(c) Context-aided Forecast (RCRPS=0.006)

Figure 37: Model: Llama-3.3-70B-Inst. The reasoning trace is evaluated as **Correct Reasoning**. The forecast changes by 91.9% and is hence evaluated as **Improvement with Context**.

D.6.4 Task: *ExplicitWithDatesAndDaysTrafficForecastTaskwithHolidaysInPredictionWindow*

Background: This series contains the road occupancy rates on a freeway in the San Francisco Bay area. The days for which the forecast is required are Thursday 2024-07-04, Friday 2024-07-05, Saturday 2024-07-06. Note that 2024-07-04 is a holiday due to Independence Day. Note that traffic on this freeway typically reduces on holidays.
Constraints: None.
Scenario: None.

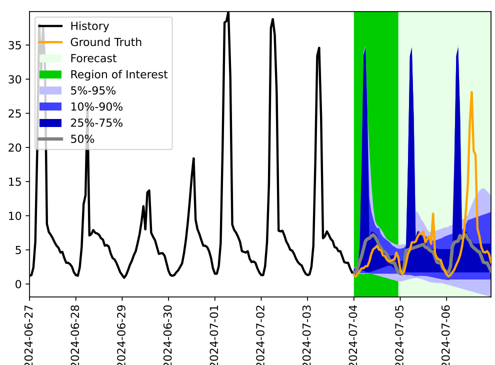
Figure 38: Context

A logical forecast can be produced by analyzing historical occupancy rates for the same weekdays and dates, especially focusing on past Independence Days and adjacent days, to account for typical holiday traffic reductions and altered travel patterns.

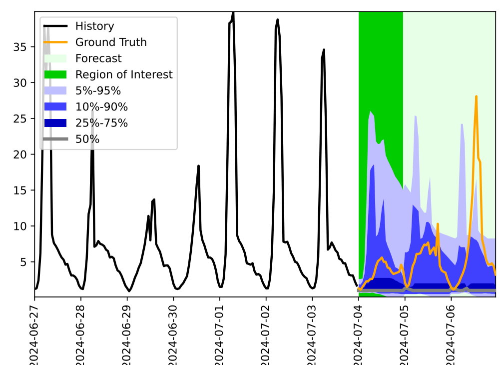
Figure 39: Gold Standard Reasoning Trace

To create an effective forecast, let's break down the context and steps needed:
1. **Understand the Context**: The given data shows daily occupancy rates for a freeway in the San Francisco Bay area over a week. We need to focus on Thursday, Friday, and Saturday forecasts for July 4th, 2024, and July 5th, 2024, noting that these dates fall on a holiday (Independence Day).
2. **Holiday Impact**: Holidays generally see a reduction in traffic volumes compared to regular days. Observing past data on similar holidays can help estimate the occupancy rate on July 4th, 2024.
3. **Pattern Analysis**: Identify if there are any seasonal patterns or trends that we can leverage. For example, weekend traffic might be slightly higher than weekdays but less than the peak on workdays.
4. **Use Historical Data**: Use the provided historical data to train our model and make predictions.
5. **Consider Known Holidays**: Since Independence Day falls on a Thursday, we know the occupancy rate will likely be lower than usual for that day.
6. **Forecast the Remaining Days**: After understanding the Independence Day pattern, we can extrapolate the occupancy rates for the remaining days in the forecast.

(a) Reasoning trace



(b) Forecast without context (RCRPS=0.135)

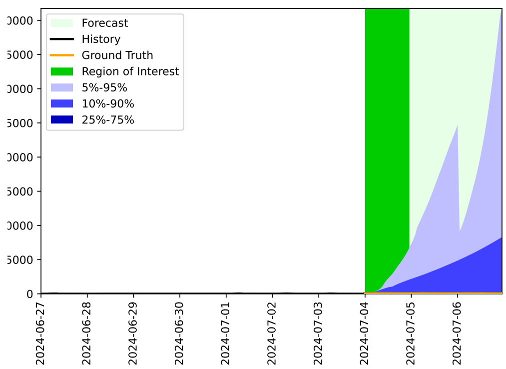


(c) Context-aided Forecast (RCRPS=0.148)

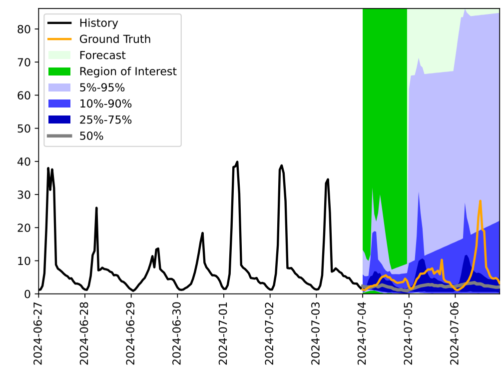
Figure 40: Model: Qwen2.5-3B-Inst. The reasoning trace is evaluated as **Correct Reasoning**. The forecast changes by -9.7% and is hence evaluated as **No Improvement with Context**.

I'll incorporate the given context into the forecasting process as follows:
1. Understanding the historical data: I'll analyze the historical time series data to identify any patterns, trends, or seasonal cycles that can be used to inform the forecast.
2. Accounting for the holiday: Given that July 4th is a holiday, which typically results in reduced traffic, I'll adjust the forecast accordingly. This may involve applying a lower growth rate or a more conservative estimate to the forecasted values.
3. Incorporating additional context: The context does not provide any specific additional information, but I'll consider the typical daily patterns in traffic, such as the morning and evening commutes.

(a) Reasoning trace



(b) Forecast without context (RCRPS=6.173)

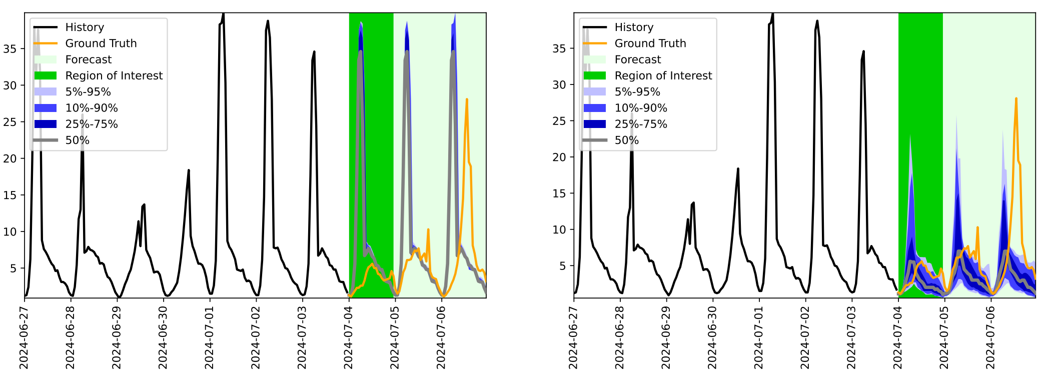


(c) Context-aided Forecast (RCRPS=0.134)

Figure 41: Model: Llama3.2-3B-Inst. The reasoning trace is evaluated as **Correct Reasoning**. The forecast changes by 97.8% and is hence evaluated as **Improvement with Context**.

To incorporate the given context effectively into our forecasting model, we need to consider the following aspects:
1. **Holiday Effect**: Since July 4th is a holiday (Independence Day), we expect lower occupancy rates compared to regular Thursdays. We can observe from the history that typical Thursday patterns show higher occupancy during morning and evening rush hours, whereas holiday patterns usually have reduced traffic throughout the day.
2. **Day of Week Patterns**: Regular days follow certain patterns. Typically, weekdays like Thursday, Friday, and Saturday (though Saturday is usually quieter) have different occupancy levels based on the time of day. For example, mornings and evenings are busier than midday or late night.
3. **Trend Analysis**: From the historical data, we notice that there is a significant drop in occupancy rates as we move towards late hours. The peak times for congestion tend to be around 7 AM to 8 AM and 5 PM to 6 PM on normal days.
Given that July 4th is a holiday, we can adjust the expected occupancy rates to be generally lower across all times of the day. For July 5th and 6th, since they are a Friday and Saturday, respectively, we expect slightly lower than usual traffic but still more than a full holiday.
Therefore, we will adjust the forecasted values by considering these patterns and the impact of the holiday.

(a) Reasoning trace



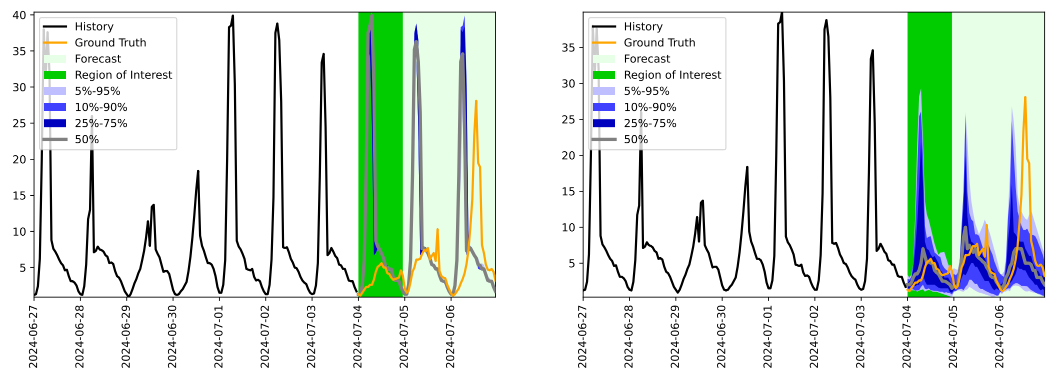
(b) Forecast without context (RCRPS=0.563)

(c) Context-aided Forecast (RCRPS=0.120)

Figure 42: Model: Qwen2.5-14B-Inst. The reasoning trace is evaluated as **Correct Reasoning**. The forecast changes by 78.7% and is hence evaluated as **Improvement with Context**.

1. **Analyzing Holiday Impact**: Since 2024-07-04 is a holiday (Independence Day), traffic is expected to be significantly lower than on a regular day. This is a key consideration for the forecast. 2. **Identifying Patterns**: From the historical data, there are clear patterns during different times of the day. For example, traffic peaks around 6 AM and drops off around 9 AM before picking up again around noon. 3. **Averaging and Adjusting for Holidays**: On holidays, we can look at the average traffic from the day before the holiday (2024-07-03) and adjust for the holiday effect, which we expect to be a substantial reduction. 4. **Carrying Forward Pattern**: For the days after the holiday (2024-07-05 and 2024-07-06), we will extend the pattern observed on 2024-07-03 while accounting for the potential easing of restrictions on weekend traffic.

(a) Reasoning trace



(b) Forecast without context (RCRPS=0.642)

(c) Context-aided Forecast (RCRPS=0.108)

Figure 43: Model: Qwen-2.5-7B-Inst. The reasoning trace is evaluated as **Correct Reasoning**. The forecast changes by 83.1% and is hence evaluated as **Improvement with Context**.

E Additional Details on CorDP

E.1 CorDP Prompt

We use the following prompt for the CorDP method, where **{base_forecasts}** are replaced by the forecasts of the quantitative forecaster in the format: (timestep1, value1), (timestep2, value2), ... (timestepN, valueN) where N is the prediction length. **{history}** is replaced by the respective numerical history for the task instance in the format (timestamp, value), **{context}** is replaced by the respective textual context for the task instance, and **((pred_time))** is replaced with the prediction timesteps.

```
I have a time series forecasting task for you.

Here is some context about the task. Make sure to factor in any background
    knowledge,
satisfy any constraints, and respect any scenarios.
<context>
{context}
</context>

Here is a historical time series in (timestamp, value) format:
<history>
{history}
</history>

And these are the forecasts of my statistical forecasting model in (timestamp,
    value) format:
<base_forecast>
{base_forecasts}
</base_forecast>

My statistical forecasting model does not support taking in context as part of
    its input. I would like you to correct its forecasts to incorporate the
    context wherever necessary, and return the corrected context-aware forecast.
Return the corrected forecast in (timestamp, value) format in between
    <corrected_forecast> and </corrected_forecast> tags.
Do not include any other information (e.g., comments) in the forecast.
```

Model	Direct Prompt (DP)	Median Corrector (Median-CorDP)			SampleWise Corrector (SampleWise-CorDP)		
		Lag-Llama	Chronos Large	arima	Lag-Llama	Chronos Large	arima
Qwen2.5-0.5B-Inst	0.339 ± 0.010	0.302 ± 0.001	0.553 ± 0.000	0.336 ± 0.001	$\mathbf{0.235 \pm 0.006}$	0.438 ± 0.014	0.272 ± 0.004
Qwen2.5-1.5B-Inst	0.317 ± 0.020	0.296 ± 0.002	0.538 ± 0.005	0.323 ± 0.002	$\mathbf{0.232 \pm 0.005}$	0.478 ± 0.007	0.278 ± 0.006
Qwen2.5-3B-Inst	0.269 ± 0.015	0.391 ± 0.004	0.420 ± 0.004	0.274 ± 0.005	$\mathbf{0.219 \pm 0.005}$	0.388 ± 0.008	0.243 ± 0.004
Qwen2.5-7B-Inst	0.285 ± 0.006	$\mathbf{0.125 \pm 0.002}$	0.198 ± 0.006	0.182 ± 0.004	0.135 ± 0.004	0.180 ± 0.006	0.146 ± 0.004
Qwen2.5-14B-Inst	$\mathbf{0.162 \pm 0.005}$	0.288 ± 0.002	0.247 ± 0.002	0.236 ± 0.005	0.206 ± 0.007	0.221 ± 0.004	0.205 ± 0.005
Qwen2.5-32B-Inst	$\mathbf{0.116 \pm 0.001}$	0.213 ± 0.002	0.156 ± 0.002	0.187 ± 0.002	0.145 ± 0.005	0.132 ± 0.002	0.137 ± 0.005
Qwen2.5-72B-Inst	$\mathbf{0.115 \pm 0.004}$	0.158 ± 0.003	0.169 ± 0.002	0.141 ± 0.003	0.138 ± 0.006	0.140 ± 0.004	0.125 ± 0.004
Llama-3.2-1B-Inst	0.336 ± 0.026	0.281 ± 0.004	0.414 ± 0.013	0.311 ± 0.002	$\mathbf{0.234 \pm 0.006}$	0.507 ± 0.003	0.269 ± 0.005
Llama-3.2-3B-Inst	0.281 ± 0.013	0.243 ± 0.003	0.368 ± 0.006	0.262 ± 0.004	$\mathbf{0.214 \pm 0.005}$	0.362 ± 0.007	0.243 ± 0.004
Llama-3-8B-Inst	0.255 ± 0.008	0.167 ± 0.005	0.189 ± 0.004	0.164 ± 0.003	$\mathbf{0.149 \pm 0.005}$	0.176 ± 0.006	0.150 ± 0.005
Llama3.3-70B-Inst	$\mathbf{0.105 \pm 0.003}$	0.211 ± 0.001	0.163 ± 0.001	0.205 ± 0.001	0.164 ± 0.005	0.126 ± 0.003	0.152 ± 0.003
Llama3.1-405B-Inst	0.126 ± 0.004	0.212 ± 0.004	0.146 ± 0.003	0.168 ± 0.003	0.131 ± 0.006	$\mathbf{0.117 \pm 0.003}$	0.144 ± 0.003
Base Quantitative Forecaster	-	0.224 ± 0.005	0.536 ± 0.003	0.272 ± 0.004	0.224 ± 0.005	0.536 ± 0.003	0.272 ± 0.004

Table 6: Results (RoI CRPS) on RoI tasks in CiK. The best-performing method with each model in every group is in **bold**.

Model	Direct Prompt (DP)	Median Corrector (Median-CorDP)			SampleWise Corrector (SampleWise-CorDP)		
		Lag-Llama	Chronos Large	arima	Lag-Llama	Chronos Large	arima
Qwen2.5-0.5B-Inst	0.129 ± 0.010	0.283 ± 0.001	0.142 ± 0.000	0.206 ± 0.001	0.211 ± 0.006	$\mathbf{0.111 \pm 0.014}$	0.159 ± 0.004
Qwen2.5-1.5B-Inst	0.224 ± 0.020	0.268 ± 0.002	0.140 ± 0.005	0.198 ± 0.002	0.193 ± 0.005	$\mathbf{0.113 \pm 0.007}$	0.160 ± 0.006
Qwen2.5-3B-Inst	0.186 ± 0.015	0.251 ± 0.004	0.129 ± 0.004	0.179 ± 0.005	0.179 ± 0.005	$\mathbf{0.114 \pm 0.008}$	0.134 ± 0.004
Qwen2.5-7B-Inst	0.164 ± 0.006	0.225 ± 0.002	0.137 ± 0.006	0.182 ± 0.004	0.167 ± 0.004	$\mathbf{0.127 \pm 0.006}$	0.146 ± 0.004
Qwen2.5-14B-Inst	$\mathbf{0.146 \pm 0.005}$	0.306 ± 0.002	0.212 ± 0.002	0.219 ± 0.005	0.210 ± 0.007	0.188 ± 0.004	0.200 ± 0.005
Qwen2.5-32B-Inst	0.140 ± 0.001	0.238 ± 0.002	0.143 ± 0.002	0.194 ± 0.002	0.164 ± 0.005	$\mathbf{0.112 \pm 0.002}$	0.131 ± 0.005
Qwen2.5-72B-Inst	$\mathbf{0.138 \pm 0.004}$	0.265 ± 0.003	0.192 ± 0.002	0.200 ± 0.003	0.181 ± 0.006	0.155 ± 0.004	0.158 ± 0.004
Llama-3.2-1B-Inst	0.248 ± 0.026	0.260 ± 0.004	0.107 ± 0.013	0.191 ± 0.002	0.191 ± 0.006	$\mathbf{0.104 \pm 0.003}$	0.159 ± 0.005
Llama-3.2-3B-Inst	0.162 ± 0.013	0.213 ± 0.003	0.116 ± 0.006	0.152 ± 0.004	0.177 ± 0.005	$\mathbf{0.107 \pm 0.007}$	0.136 ± 0.004
Llama-3-8B-Inst	$\mathbf{0.163 \pm 0.008}$	0.257 ± 0.005	0.238 ± 0.004	0.208 ± 0.003	0.232 ± 0.005	0.198 ± 0.006	0.189 ± 0.005
Llama3.3-70B-Inst	0.182 ± 0.003	0.277 ± 0.001	0.157 ± 0.001	0.194 ± 0.001	0.205 ± 0.005	$\mathbf{0.132 \pm 0.003}$	0.154 ± 0.003
Llama3.1-405B-Inst	0.150 ± 0.004	0.248 ± 0.004	0.170 ± 0.003	0.174 ± 0.003	0.163 ± 0.006	$\mathbf{0.133 \pm 0.003}$	0.141 ± 0.003
Base Quantitative Forecaster	-	0.202 ± 0.005	0.115 ± 0.003	0.159 ± 0.004	0.202 ± 0.005	0.115 ± 0.003	0.159 ± 0.004

Table 7: Results (non-RoI CRPS) on RoI tasks in CiK. The best-performing method with each model in every group is in **bold**.

E.2 Results on various groups of tasks

We now look into results aggregated across the various kinds of tasks in the CiK benchmark: Table 6, Table 7 showcases performance of methods within and outside the region of interest (RoI) respectively for tasks that have an RoI, Table 8 shows performance across tasks where the entire prediction window is the RoI, and Table 9 shows constraint RCRPS across tasks with constraints. We find that SampleWise-CorDP has an advantage on tasks with an RoI, achieving the best performance in most models, both within and outside the RoI. Median-CorDP however has a clear advantage on tasks where the shape of the entire forecast is influenced by the context, which make up most of the benchmark, achieving the best performance in half the models, and trailing closely with DP in the other. These results also indicate that DP methods are still consistently strong in tasks where the entire prediction is influenced

Model	Direct Prompt (DP)	Median Corrector (Median-CorDP)			SampleWise Corrector (SampleWise-CorDP)		
		Lag-Llama	Chronos Large	arima	Lag-Llama	Chronos Large	arima
Qwen2.5-0.5B-Inst	0.836 ± 0.046	0.864 ± 0.003	1.110 ± 0.006	1.094 ± 0.090	0.679 ± 0.013	0.895 ± 0.127	0.953 ± 0.092
Qwen2.5-1.5B-Inst	0.851 ± 0.026	**0.525 ± 0.021**	0.672 ± 0.005	0.969 ± 0.011	0.733 ± 0.030	0.595 ± 0.007	1.059 ± 0.021
Qwen2.5-3B-Inst	0.558 ± 0.027	0.606 ± 0.008	0.638 ± 0.006	0.849 ± 0.014	**0.533 ± 0.048**	0.587 ± 0.007	0.731 ± 0.053
Qwen2.5-7B-Inst	**0.521 ± 0.009**	0.584 ± 0.006	0.964 ± 0.013	0.939 ± 0.013	0.538 ± 0.011	0.571 ± 0.034	0.808 ± 0.019
Qwen2.5-14B-Inst	**0.310 ± 0.010**	0.328 ± 0.004	0.406 ± 0.009	0.556 ± 0.007	0.470 ± 0.009	0.551 ± 0.009	0.654 ± 0.015
Qwen2.5-32B-Inst	0.580 ± 0.013	**0.263 ± 0.007**	0.355 ± 0.009	0.423 ± 0.013	0.416 ± 0.008	0.486 ± 0.011	0.604 ± 0.014
Qwen2.5-72B-Inst	**0.253 ± 0.015**	0.392 ± 0.014	0.479 ± 0.017	0.603 ± 0.015	0.320 ± 0.016	0.441 ± 0.016	0.552 ± 0.017
Llama-3.2-1B-Inst	**0.467 ± 0.041**	0.477 ± 0.007	0.687 ± 0.008	0.857 ± 0.030	0.765 ± 0.014	0.857 ± 0.008	0.983 ± 0.025
Llama-3.2-3B-Inst	1.004 ± 0.040	**0.422 ± 0.018**	0.600 ± 0.014	0.821 ± 0.037	0.722 ± 0.043	0.551 ± 0.012	0.985 ± 0.052
Llama-3-8B-Inst	0.771 ± 0.043	**0.385 ± 0.006**	0.615 ± 0.008	0.833 ± 0.007	0.586 ± 0.015	0.561 ± 0.006	0.953 ± 0.016
Llama3.3-70B-Inst	0.289 ± 0.011	0.306 ± 0.004	0.313 ± 0.006	0.456 ± 0.010	**0.249 ± 0.006**	0.273 ± 0.006	0.419 ± 0.011
Llama3.1-405B-Inst	**0.196 ± 0.005**	0.310 ± 0.014	0.272 ± 0.006	0.316 ± 0.012	0.235 ± 0.009	0.241 ± 0.006	0.288 ± 0.013
Base Quantitative Forecaster	-	0.497 ± 0.018	0.605 ± 0.006	0.921 ± 0.023	0.497 ± 0.018	0.605 ± 0.006	0.921 ± 0.023

Table 8: Results (RCRPS) on tasks with a full RoI in CiK. The best-performing method with each model in every group is in **bold**.

Model	Direct Prompt (DP)	Median Corrector (Median-CorDP)			SampleWise Corrector (SampleWise-CorDP)		
		Lag-Llama	Chronos Large	arima	Lag-Llama	Chronos Large	arima
Qwen2.5-0.5B-Inst	0.243 ± 0.103	**0.116 ± 0.007**	0.501 ± 0.008	0.675 ± 0.025	0.236 ± 0.028	0.861 ± 0.204	0.716 ± 0.044
Qwen2.5-1.5B-Inst	0.706 ± 0.147	**0.185 ± 0.047**	0.488 ± 0.008	0.680 ± 0.022	0.794 ± 0.065	0.485 ± 0.010	1.185 ± 0.043
Qwen2.5-3B-Inst	0.234 ± 0.056	0.483 ± 0.008	0.478 ± 0.005	0.469 ± 0.024	0.418 ± 0.107	0.474 ± 0.005	0.422 ± 0.118
Qwen2.5-7B-Inst	0.470 ± 0.078	0.507 ± 0.009	0.947 ± 0.004	0.547 ± 0.026	0.523 ± 0.015	**0.146 ± 0.065**	0.537 ± 0.036
Qwen2.5-14B-Inst	0.039 ± 0.015	0.001 ± 0.003	**0.000 ± 0.005**	0.051 ± 0.009	0.457 ± 0.015	0.455 ± 0.010	0.466 ± 0.030
Qwen2.5-32B-Inst	0.479 ± 0.019	0.001 ± 0.009	**0.000 ± 0.005**	0.000 ± 0.027	0.758 ± 0.012	0.455 ± 0.008	0.455 ± 0.028
Qwen2.5-72B-Inst	0.032 ± 0.028	0.304 ± 0.006	**0.000 ± 0.006**	0.003 ± 0.007	0.004 ± 0.008	0.000 ± 0.008	0.001 ± 0.024
Llama-3.2-1B-Inst	0.275 ± 0.092	**0.084 ± 0.011**	0.482 ± 0.015	0.499 ± 0.068	0.905 ± 0.027	0.924 ± 0.013	1.168 ± 0.053
Llama-3.2-3B-Inst	1.030 ± 0.090	**0.112 ± 0.032**	0.519 ± 0.018	0.502 ± 0.081	0.884 ± 0.091	0.487 ± 0.016	1.003 ± 0.116
Llama-3-8B-Inst	0.169 ± 0.172	**0.061 ± 0.011**	0.481 ± 0.015	0.438 ± 0.034	0.609 ± 0.029	0.476 ± 0.012	0.943 ± 0.033
Llama3.3-70B-Inst	**0.000 ± 0.024**	0.000 ± 0.003	0.001 ± 0.007	0.000 ± 0.022	0.002 ± 0.010	0.000 ± 0.006	0.000 ± 0.022
Llama3.1-405B-Inst	0.004 ± 0.009	0.060 ± 0.031	0.303 ± 0.008	0.006 ± 0.025	0.042 ± 0.016	**0.000 ± 0.009**	0.228 ± 0.027
Base Quantitative Forecaster	-	0.204 ± 0.037	0.487 ± 0.010	0.843 ± 0.050	0.204 ± 0.037	0.487 ± 0.010	0.843 ± 0.050

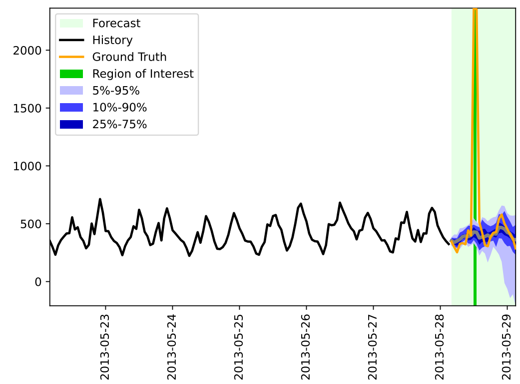
Table 9: Results (Constraint violation CRPS) on tasks with constraints. The best-performing method with each model in every group is in **bold**.

by the context. Median-CorDP overwhelmingly outperforms DP and bags the best performance in tasks with constraints, sometimes achieving perfect performance with large models. This shows that when choosing between CorDP methods, the kind of tasks that will be encountered is an important factor to consider.

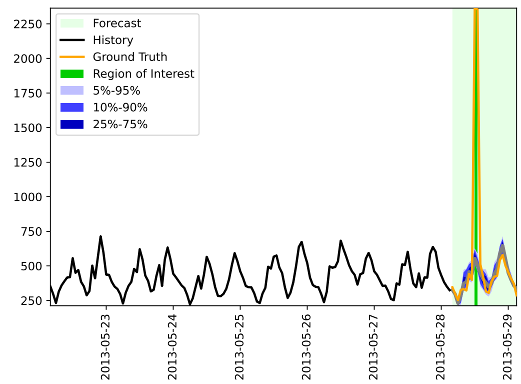
E.3 Example Forecasts

E.3.1 Task: *ElectricityIncreaseInPredictionWithSplitContext*

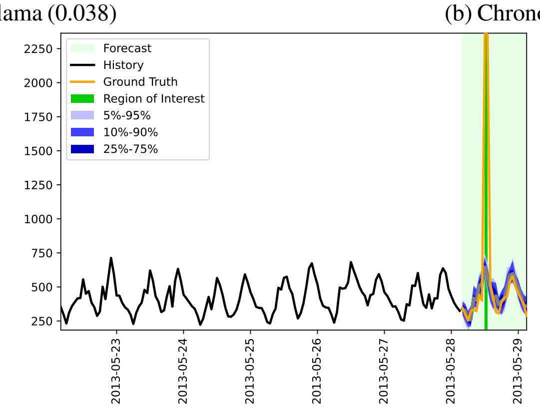
Context:
Background: This is the electricity consumption recorded in Kilowatt (kW) in city A.
Constraints: None
Scenario: Suppose that there is a heat wave in city A from 2013-05-28 12:00:00 for 2 hours, which would typically lead to excessive use of air conditioning, and 10 times the usual electricity being consumed. But in this case, residents sought to conserve energy and used lesser air conditioning, resulting in excessive usage of only 4 times the usual electricity.



(a) Lag-Llama (0.038)

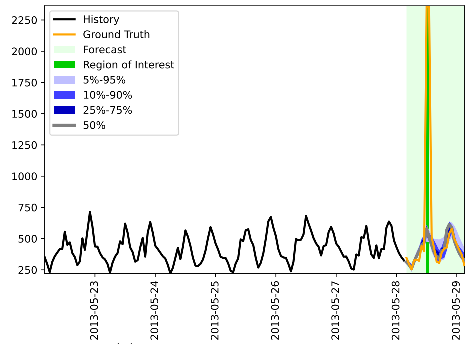


(b) Chronos (0.036)

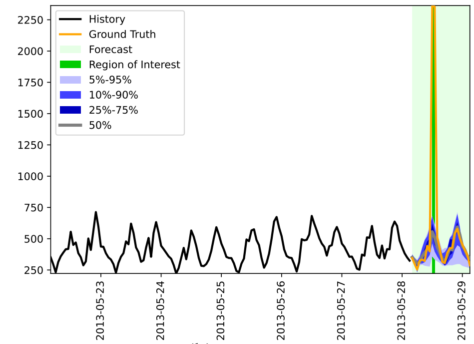


(c) ARIMA (0.035)

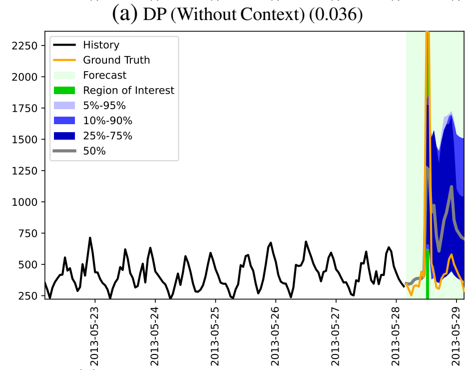
Figure 44: Forecasts of Lag-Llama, Chronos, and ARIMA on the *ElectricityIncreaseInPredictionWithSplitContext* task (with RCRPS in brackets)



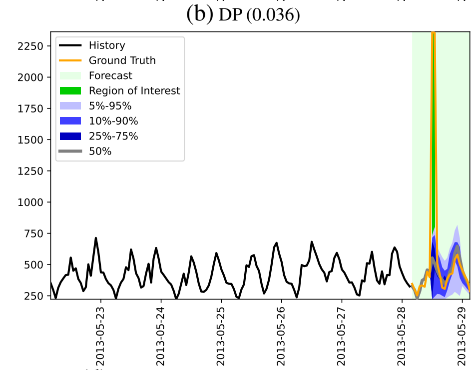
(a) DP (Without Context) (0.036)



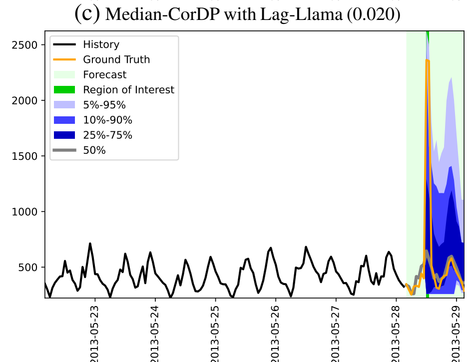
(b) DP (0.036)



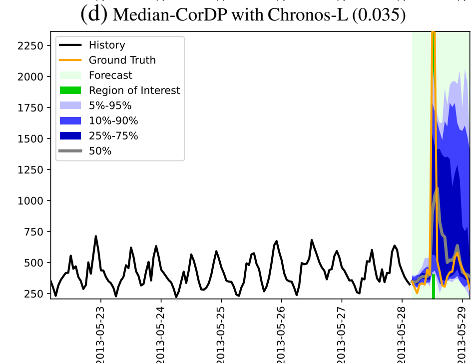
(c) Median-CorDP with Lag-Llama (0.020)



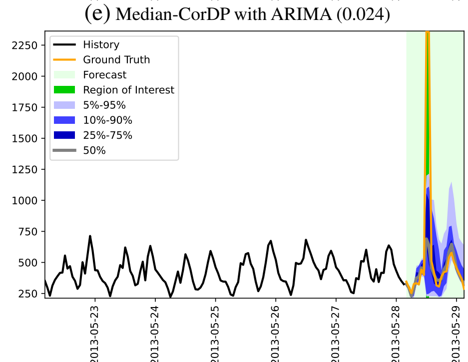
(d) Median-CorDP with Chronos-L (0.035)



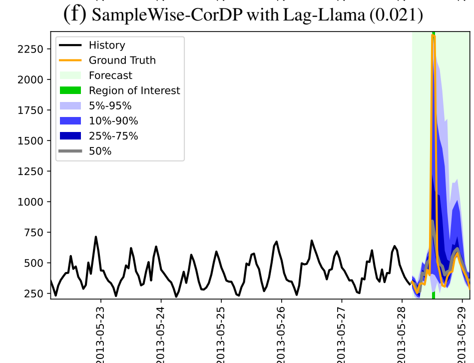
(e) Median-CorDP with ARIMA (0.024)



(f) SampleWise-CorDP with Lag-Llama (0.021)

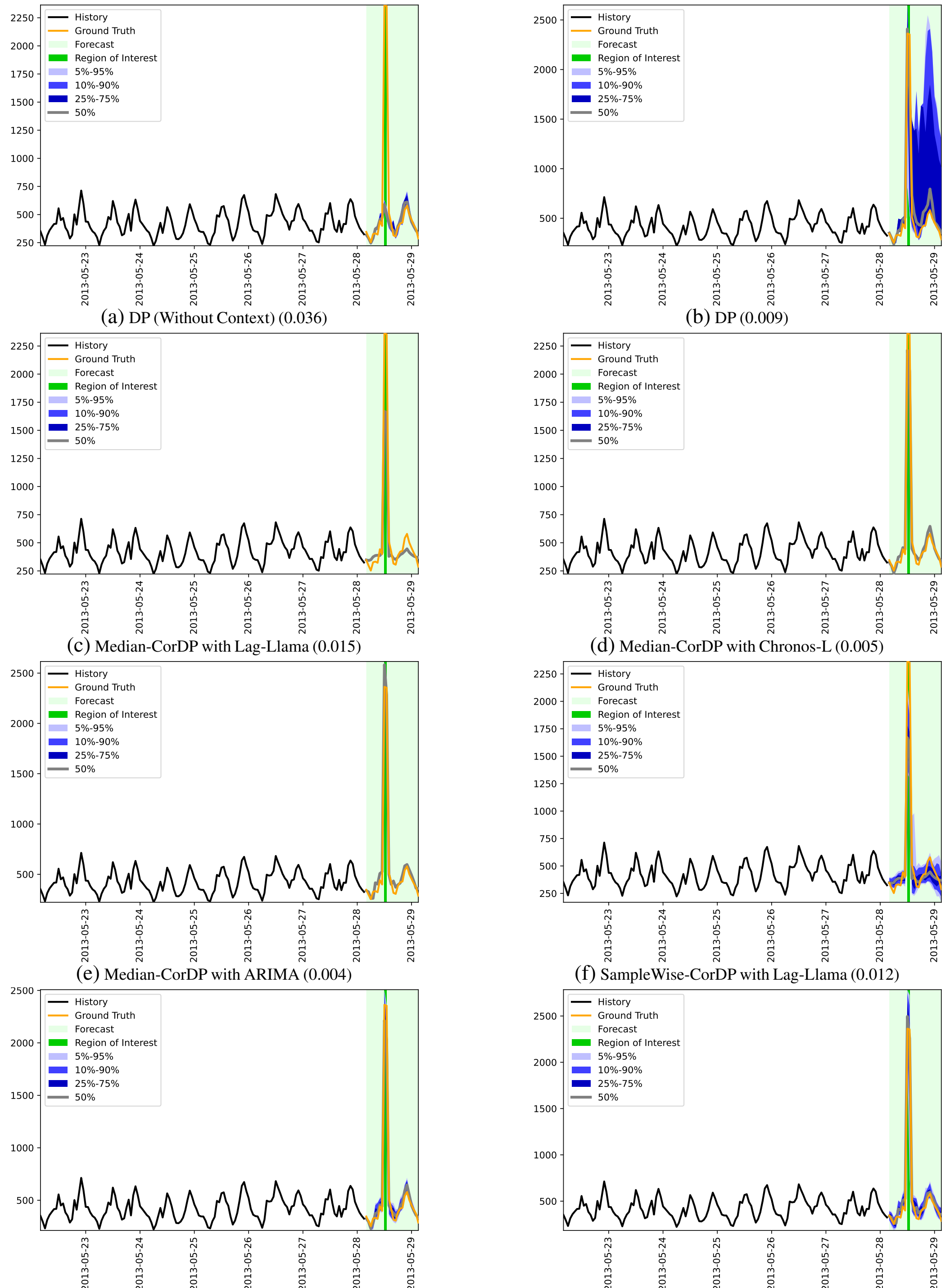


(g) SampleWise-CorDP with Chronos-L (0.029)



(h) SampleWise-CorDP with ARIMA (0.021)

Figure 45: Forecasts of model Qwen2.5-7B-Inst on task *ElectricityIncreaseInPredictionWithSplitContext* (with RCRPS in brackets)



(a) DP (Without Context) (0.036)

(b) DP (0.009)

(c) Median-CorDP with Lag-Llama (0.015)

(d) Median-CorDP with Chronos-L (0.005)

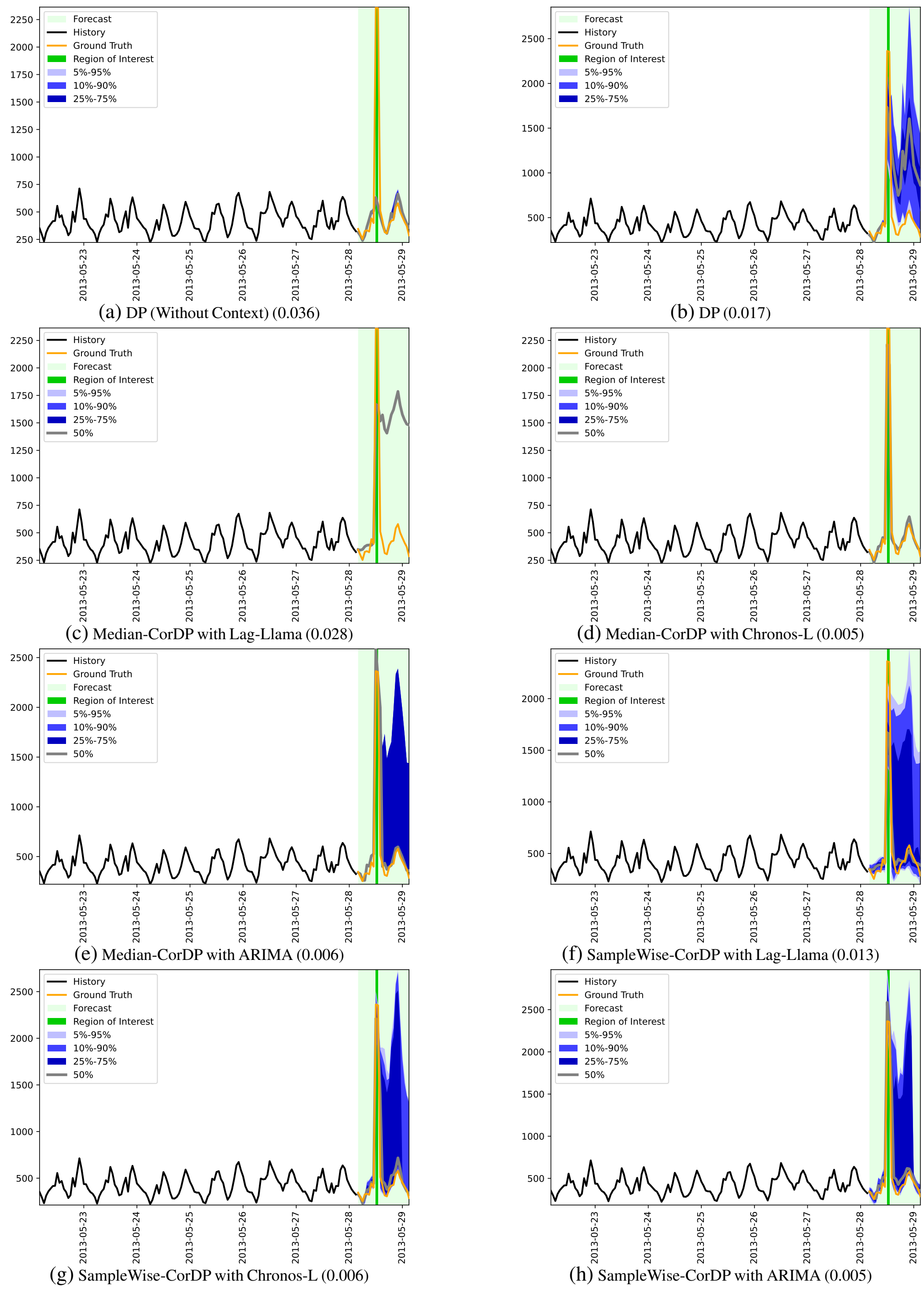
(e) Median-CorDP with ARIMA (0.004)

(f) SampleWise-CorDP with Lag-Llama (0.012)

(g) SampleWise-CorDP with Chronos-L (0.003)

(h) SampleWise-CorDP with ARIMA (0.002)

Figure 46: Forecasts of model Qwen2.5-32B-Inst on task *ElectricityIncreaseInPredictionWithSplitContext* (with RCRPS in brackets)



(a) DP (Without Context) (0.036)

(b) DP (0.017)

(c) Median-CorDP with Lag-Llama (0.028)

(d) Median-CorDP with Chronos-L (0.005)

(e) Median-CorDP with ARIMA (0.006)

(f) SampleWise-CorDP with Lag-Llama (0.013)

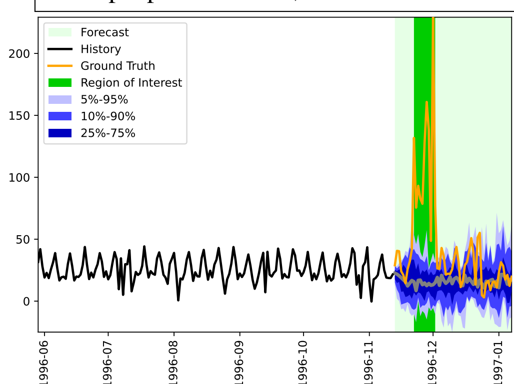
(g) SampleWise-CorDP with Chronos-L (0.006)

(h) SampleWise-CorDP with ARIMA (0.005)

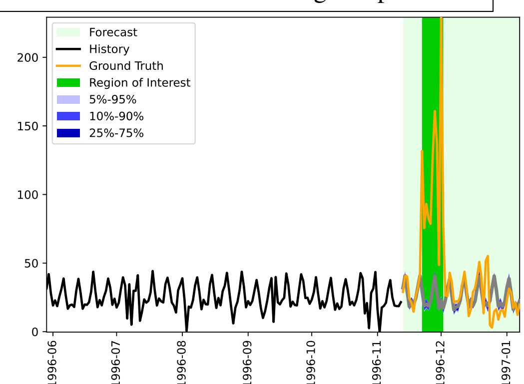
Figure 47: Forecasts of model Llama3.3-70B-Inst on task *ElectricityIncreaseInPredictionWithSplitContext* (with RCRPS in brackets)

E.3.2 Task: *IncreasedWithdrawalScenario*

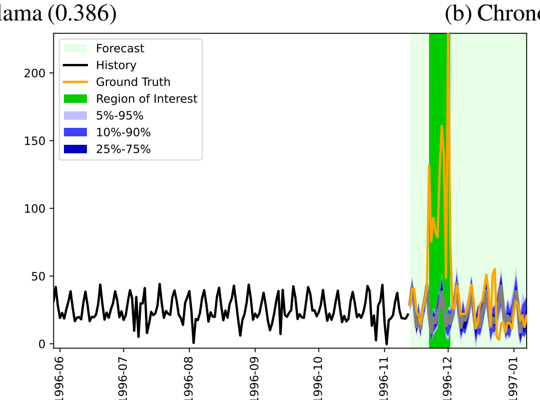
Context:
Background: This is the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England.
Constraints: None
Scenario: Suppose that there is a carnival from 1996-11-22 00:00:00, for 11 days leading to more people in the area, and 4 times the number of usual withdrawals during that period.



(a) Lag-Llama (0.386)

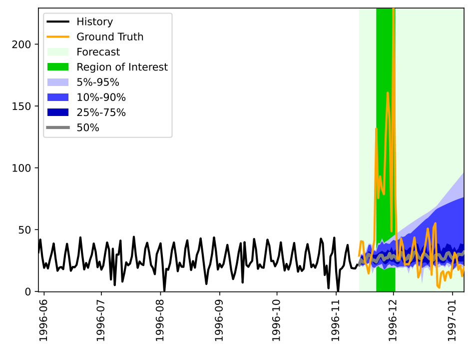


(b) Chronos (0.379)

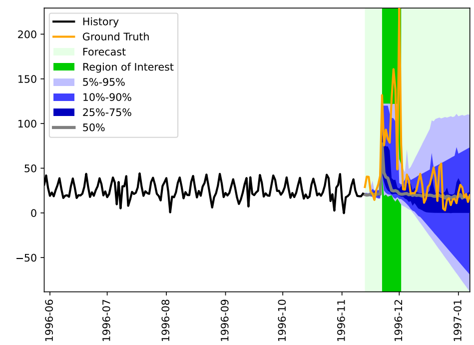


(c) ARIMA (0.380)

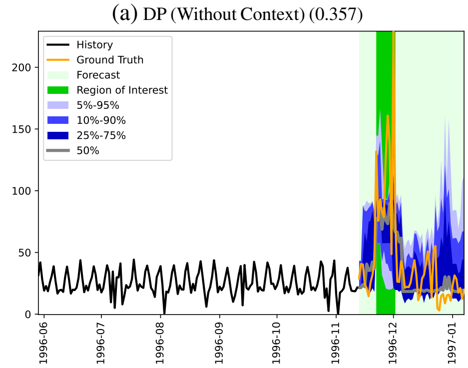
Figure 48: Forecasts of Lag-Llama, Chronos, and ARIMA on the *IncreasedWithdrawalScenario* task (with RCRPS in brackets)



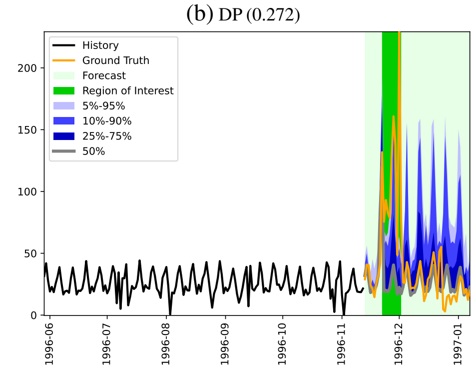
(a) DP (Without Context) (0.357)



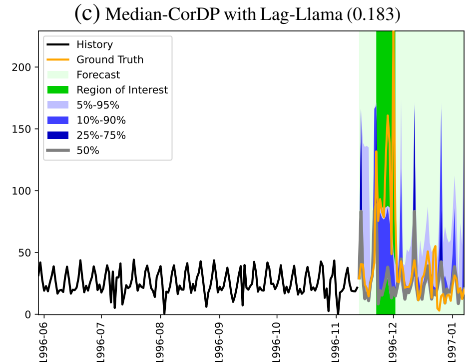
(b) DP (0.272)



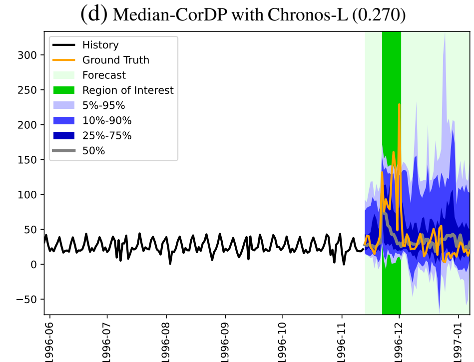
(c) Median-CorDP with Lag-Llama (0.183)



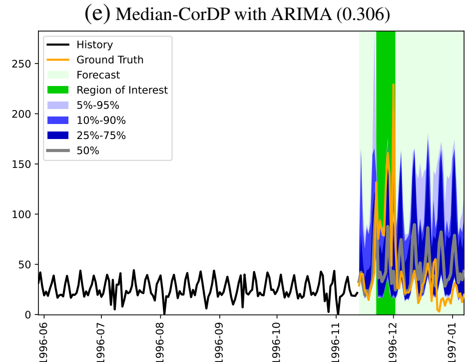
(d) Median-CorDP with Chronos-L (0.270)



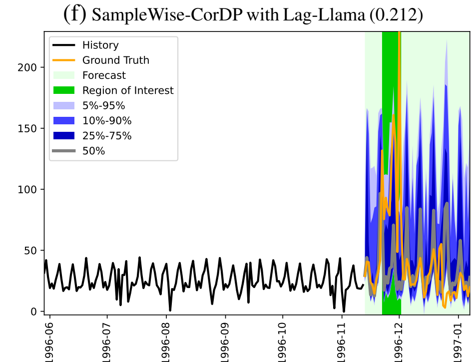
(e) Median-CorDP with ARIMA (0.306)



(f) SampleWise-CorDP with Lag-Llama (0.212)

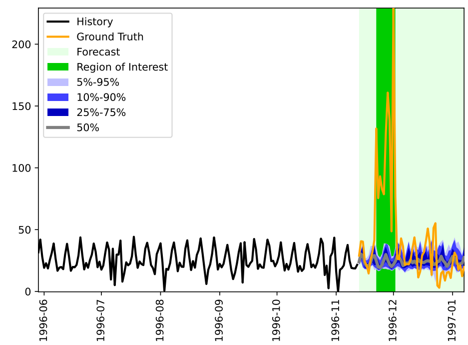


(g) SampleWise-CorDP with Chronos-L (0.230)

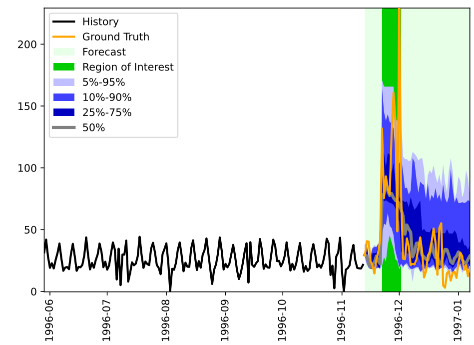


(h) SampleWise-CorDP with ARIMA (0.235)

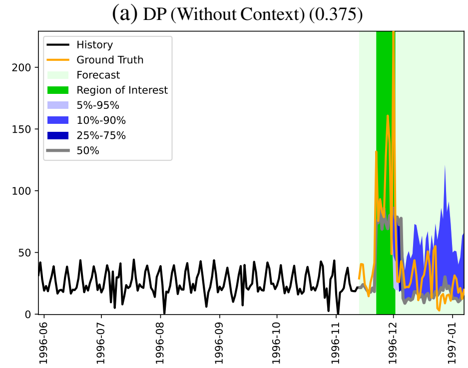
Figure 49: Forecasts of model Qwen2.5-7B-Inst on task *IncreasedWithdrawalScenario* (with RCRPS in brackets)



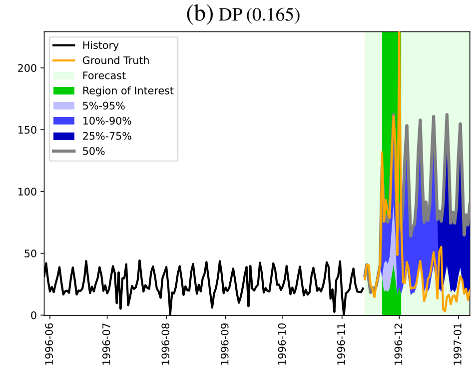
(a) DP (Without Context) (0.375)



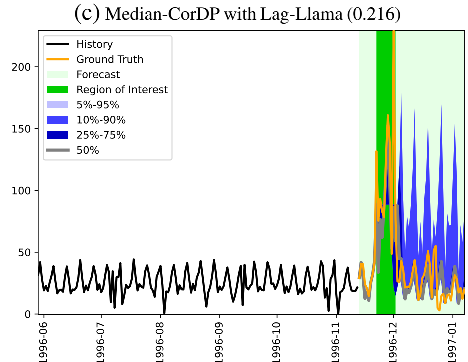
(b) DP (0.165)



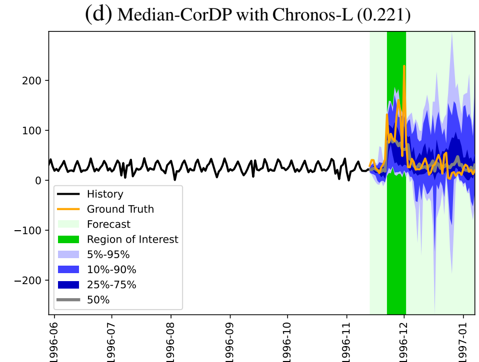
(c) Median-CorDP with Lag-Llama (0.216)



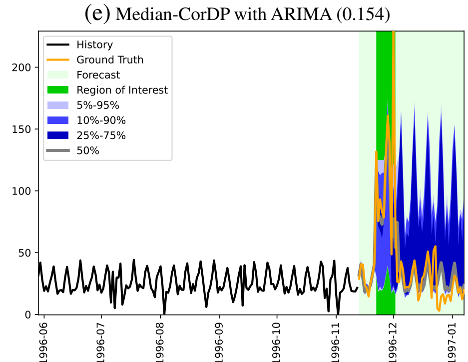
(d) Median-CorDP with Chronos-L (0.221)



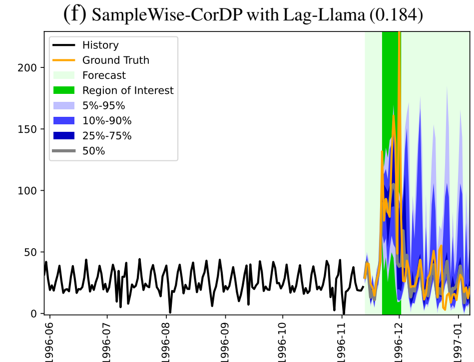
(e) Median-CorDP with ARIMA (0.154)



(f) SampleWise-CorDP with Lag-Llama (0.184)



(g) SampleWise-CorDP with Chronos-L (0.136)

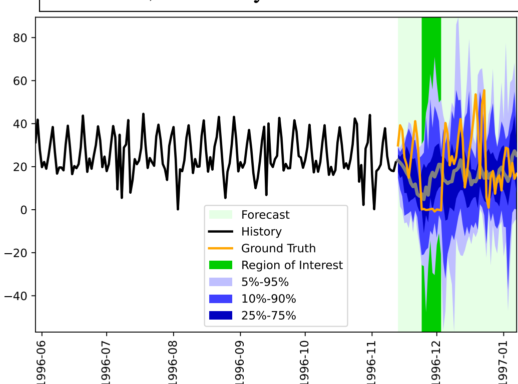


(h) SampleWise-CorDP with ARIMA (0.145)

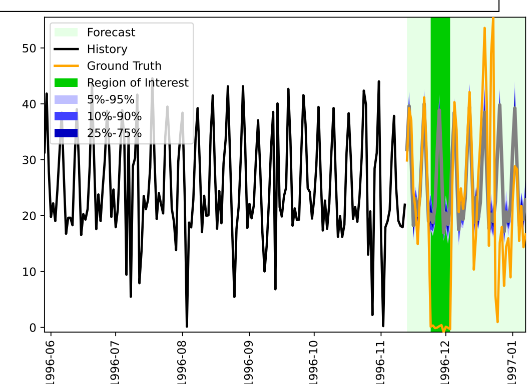
Figure 50: Forecasts of model Qwen2.5-32B-Inst on task *IncreasedWithdrawalScenario* (with RCRPS in brackets)

E.3.3 Task: *ATMBuildingClosed*

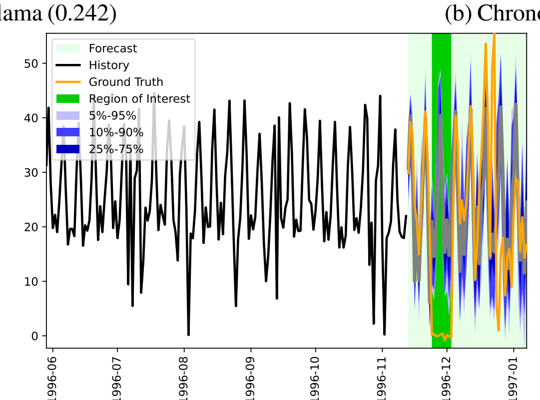
Context:
Background: This is the number of cash withdrawals from an automated teller machine (ATM) in an arbitrary location in England.
Constraints: None
Scenario: Consider that the building which contains the ATM is closed from 1996-11-24 00:00:00, for 10 days.



(a) Lag-Llama (0.242)

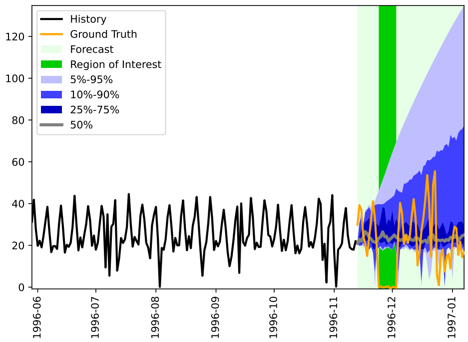


(b) Chronos (0.398)

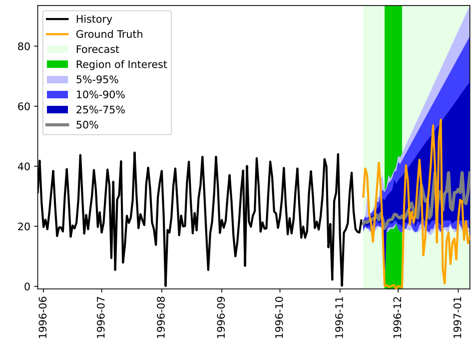


(c) ARIMA (0.337)

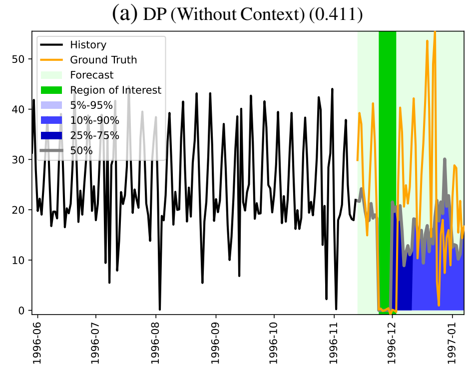
Figure 51: Forecasts of Lag-Llama, Chronos, and ARIMA on the *ATMBuildingClosedTask* task (with RCRPS in brackets)



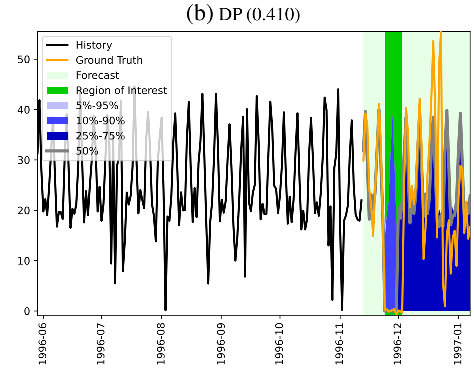
(a) DP (Without Context) (0.411)



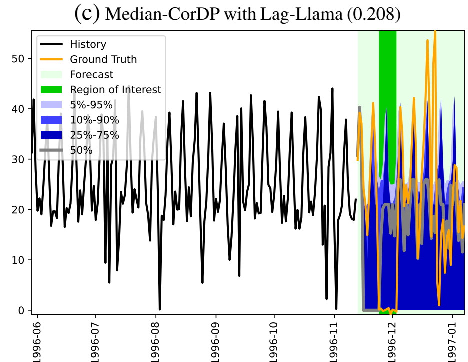
(b) DP (0.410)



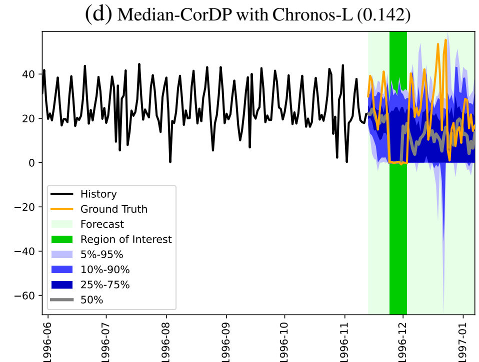
(c) Median-CorDP with Lag-Llama (0.208)



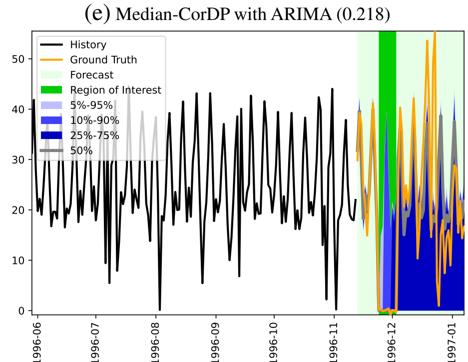
(d) Median-CorDP with Chronos-L (0.142)



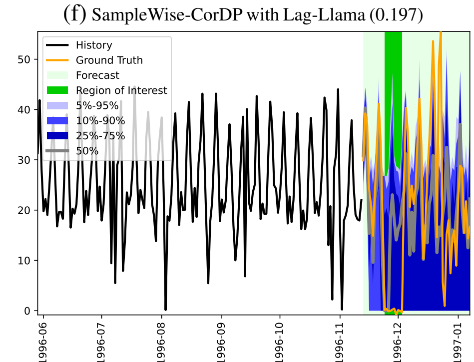
(e) Median-CorDP with ARIMA (0.218)



(f) SampleWise-CorDP with Lag-Llama (0.197)

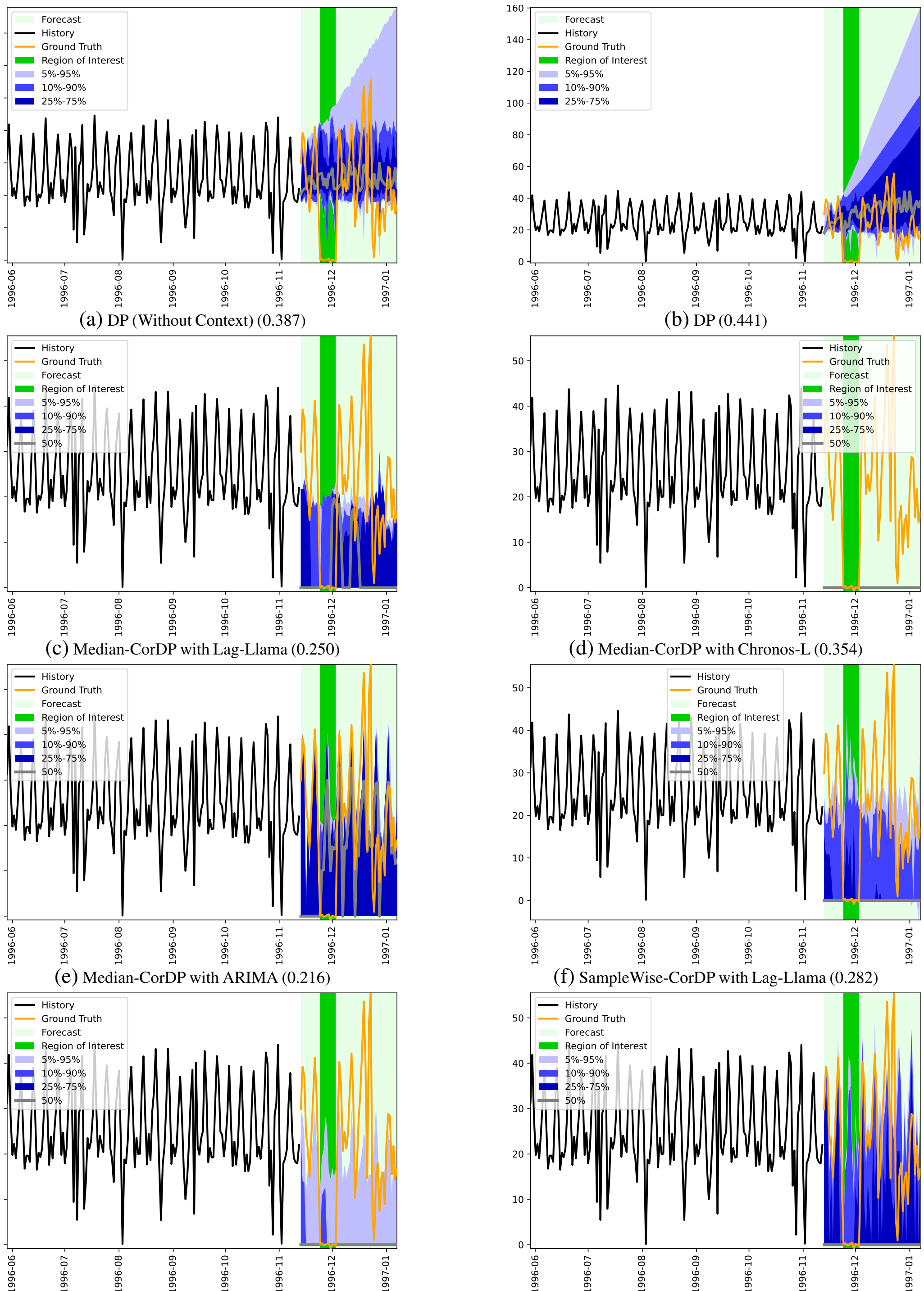


(g) SampleWise-CorDP with Chronos-L (0.115)



(h) SampleWise-CorDP with ARIMA (0.178)

Figure 52: Forecasts of model Qwen2.5-7B-Inst on task *ATMBuildingClosedTask* (with RCRPS in brackets)



(a) DP (Without Context) (0.387)

(b) DP (0.441)

(c) Median-CorDP with Lag-Llama (0.250)

(d) Median-CorDP with Chronos-L (0.354)

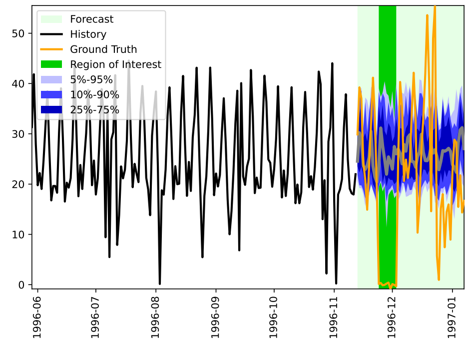
(e) Median-CorDP with ARIMA (0.216)

(f) SampleWise-CorDP with Lag-Llama (0.282)

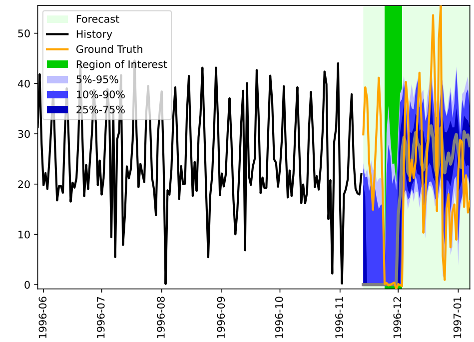
(g) SampleWise-CorDP with Chronos-L (0.306)

(h) SampleWise-CorDP with ARIMA (0.220)

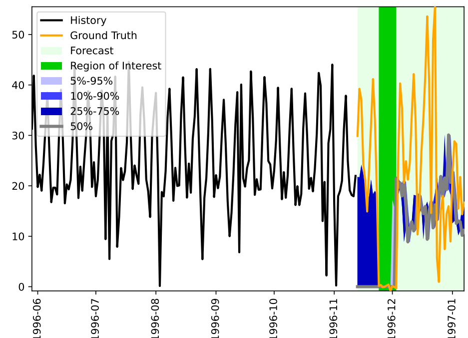
Figure 53: Forecasts of model Llama3-8B-Inst on task *ATMBuildingClosed* (with RCRPS in brackets)



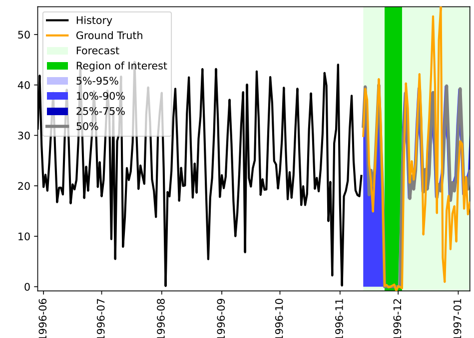
(a) DP (Without Context) (0.410)



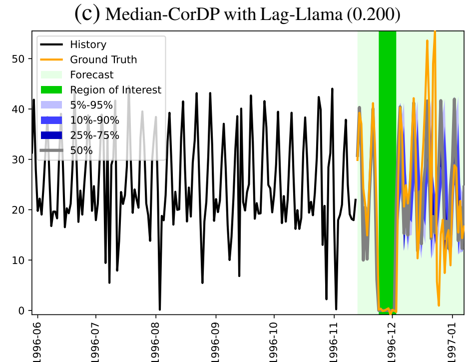
(b) DP (0.195)



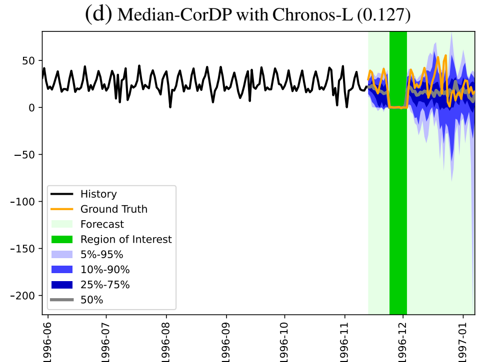
(c) Median-CorDP with Lag-Llama (0.200)



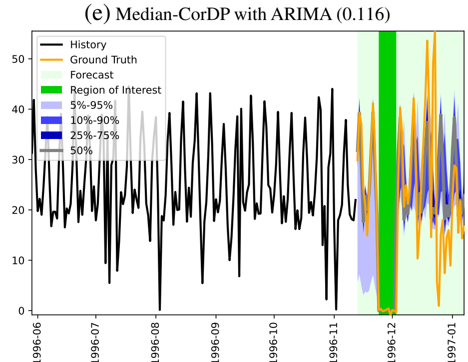
(d) Median-CorDP with Chronos-L (0.127)



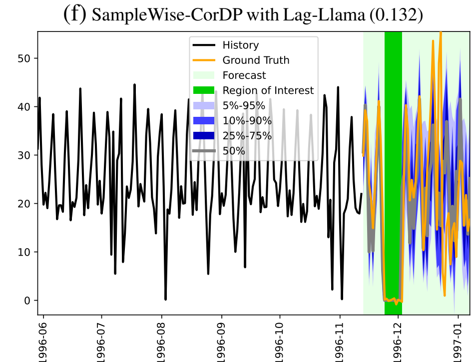
(e) Median-CorDP with ARIMA (0.116)



(f) SampleWise-CorDP with Lag-Llama (0.132)



(g) SampleWise-CorDP with Chronos-L (0.104)

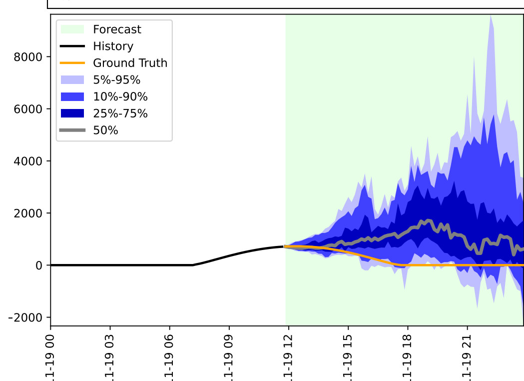


(h) SampleWise-CorDP with ARIMA (0.102)

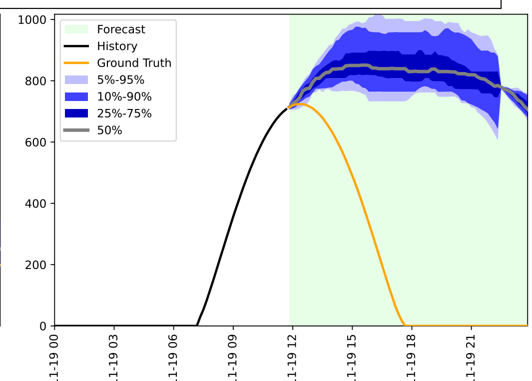
Figure 54: Forecasts of model Llama3.1-405B-Inst on task *ATMBuildingClosed* (with RCRPS in brackets)

E.3.4 Task: *ZenithInfoHalfDaySolarForecastTask*

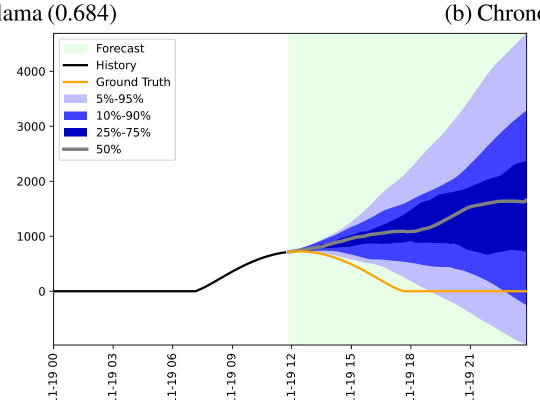
Context:
Background: This series contains the amount of sunlight (in Watts per squared meter) arriving on a horizontal surface, for a location in Florida, United States. Over the previous 90 days, the maximum sunlight happened on average at 12:25:33.
Constraints: None
Scenario: None



(a) Lag-Llama (0.684)

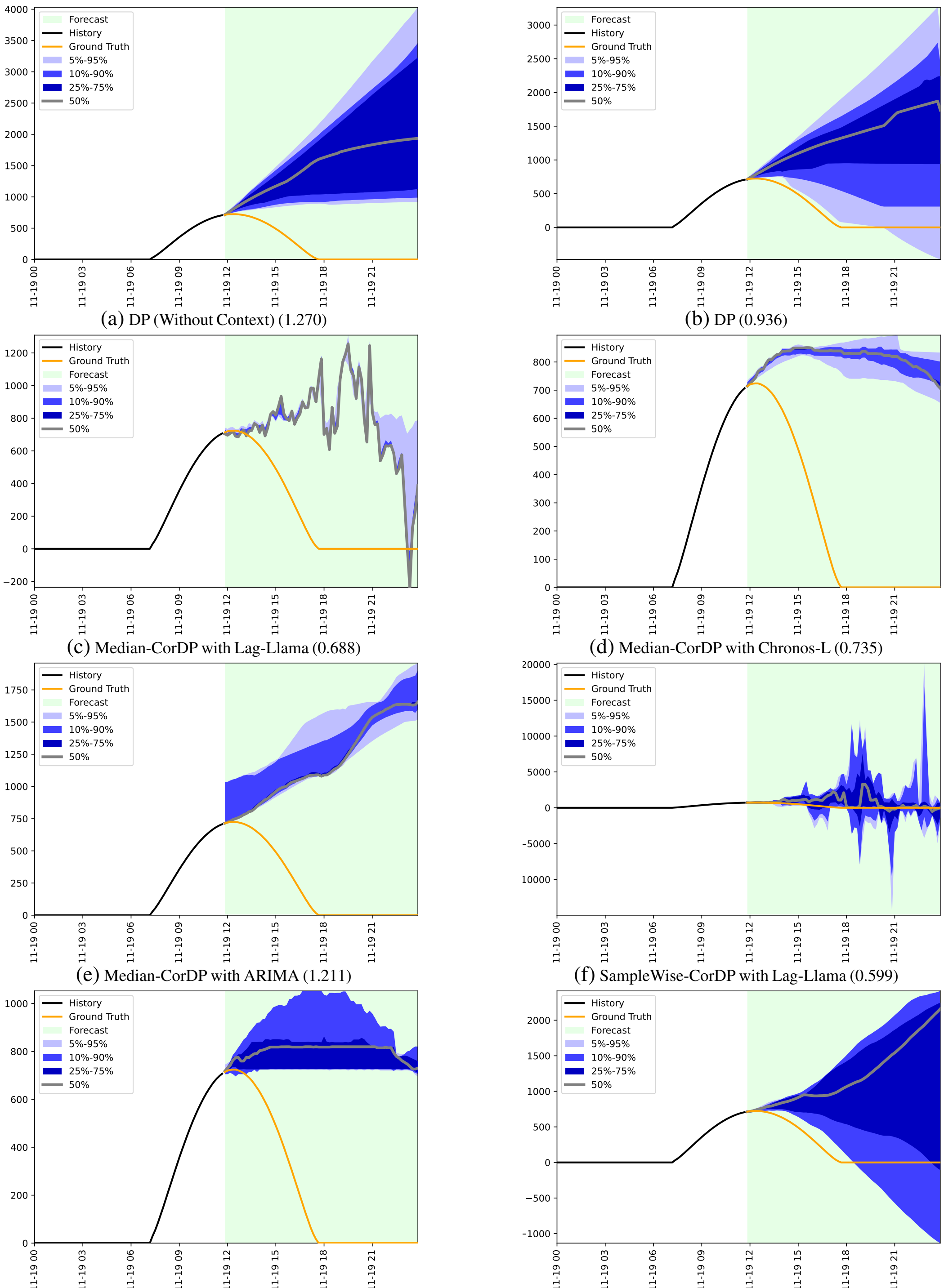


(b) Chronos (0.727)



(c) ARIMA (0.804)

Figure 55: Forecasts of Lag-Llama, Chronos, and ARIMA on the *ZenithInfoHalfDaySolarForecastTask* task (with RCRPS in brackets)



(a) DP (Without Context) (1.270)

(b) DP (0.936)

(c) Median-CorDP with Lag-Llama (0.688)

(d) Median-CorDP with Chronos-L (0.735)

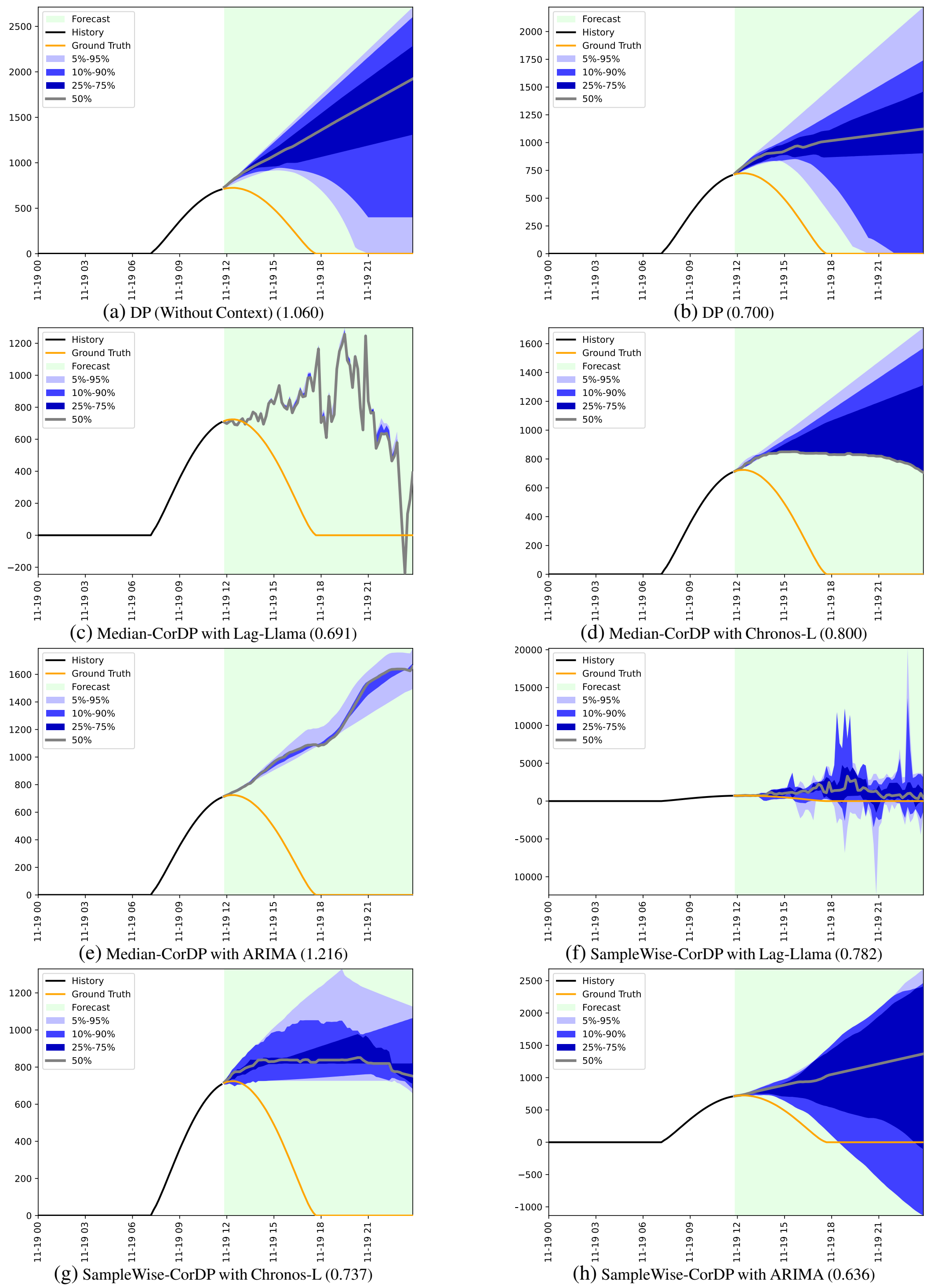
(e) Median-CorDP with ARIMA (1.211)

(f) SampleWise-CorDP with Lag-Llama (0.599)

(g) SampleWise-CorDP with Chronos-L (0.701)

(h) SampleWise-CorDP with ARIMA (0.651)

Figure 56: Forecasts of model Llama3.2-3B-Inst on task *ZenithInfoHalfDaySolarForecastTask* (with RCRPS in brackets)



(a) DP (Without Context) (1.060)

(b) DP (0.700)

(c) Median-CorDP with Lag-Llama (0.691)

(d) Median-CorDP with Chronos-L (0.800)

(e) Median-CorDP with ARIMA (1.216)

(f) SampleWise-CorDP with Lag-Llama (0.782)

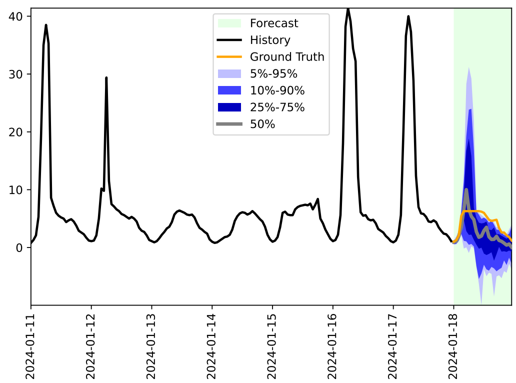
(g) SampleWise-CorDP with Chronos-L (0.737)

(h) SampleWise-CorDP with ARIMA (0.636)

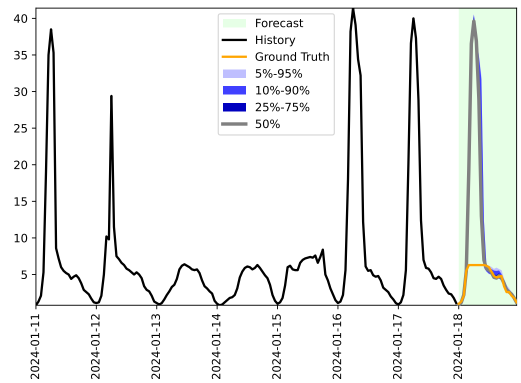
Figure 57: Forecasts of model Llama3-8B-Inst on task *ZenithInfoHalfDaySolarForecastTask* (with RCRPS in brackets)

E.3.5 Task: *BoundedPredConstraintsBasedOnPredQuantilesTask*

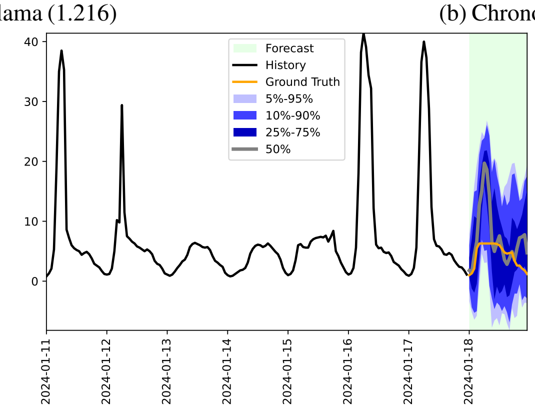
Context:
Background: None
Constraints: Suppose that in the forecast, the values are bounded above by 6.29.
Scenario: None



(a) Lag-Llama (1.216)

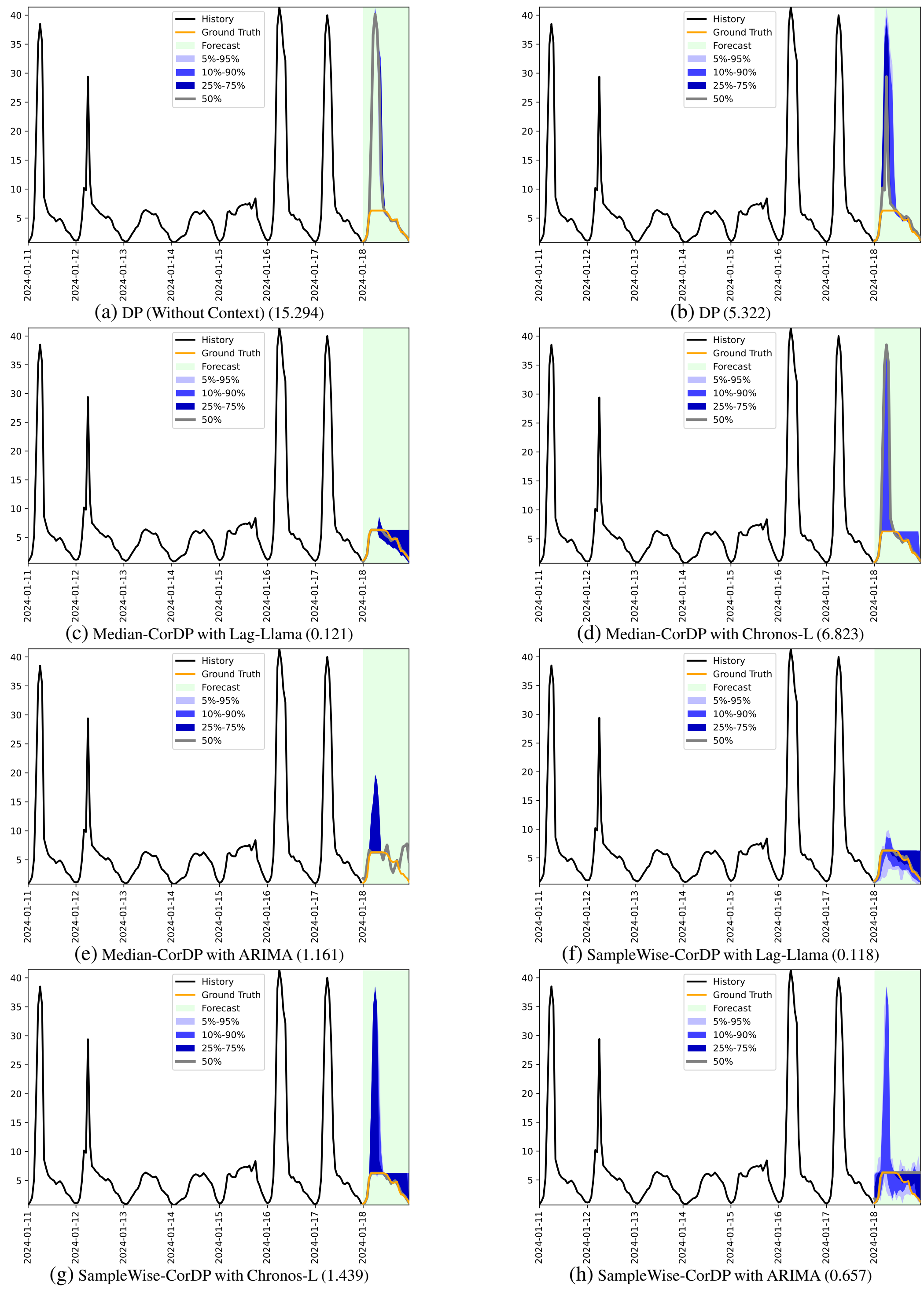


(b) Chronos (15.311)



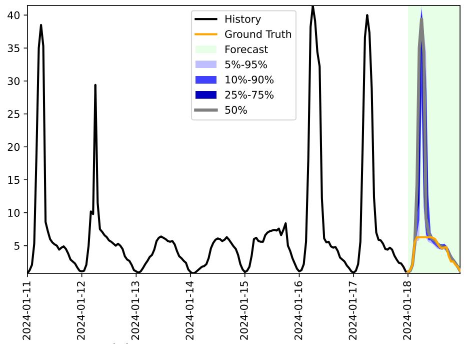
(c) ARIMA (8.232)

Figure 58: Forecasts of Lag-Llama, Chronos, and ARIMA on the *BoundedPredConstraintsBasedOnPredQuantilesTask* task (with RCRPS in brackets)

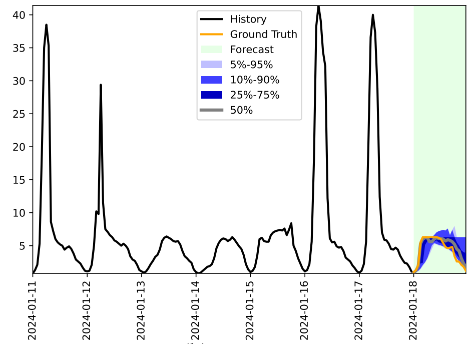


(a) DP (Without Context) (15.294)
(b) DP (5.322)
(c) Median-CorDP with Lag-Llama (0.121)
(d) Median-CorDP with Chronos-L (6.823)
(e) Median-CorDP with ARIMA (1.161)
(f) SampleWise-CorDP with Lag-Llama (0.118)
(g) SampleWise-CorDP with Chronos-L (1.439)
(h) SampleWise-CorDP with ARIMA (0.657)

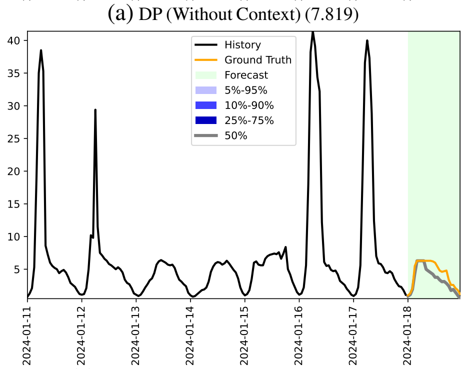
Figure 59: Forecasts of model Qwen2.5-7B-Inst on task *BoundedPredConstraintsBasedOnPredQuantilesTask* (with RCRPS in brackets)



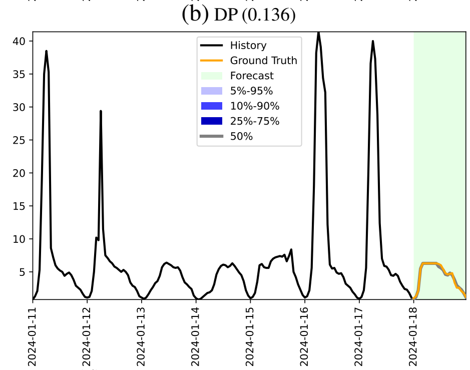
(a) DP (Without Context) (7.819)



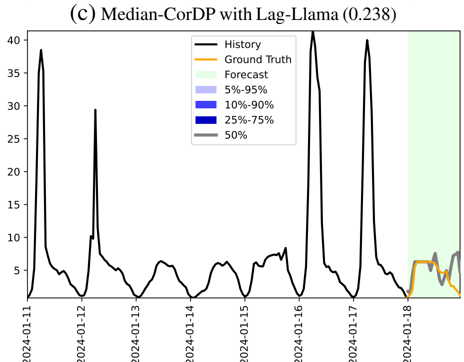
(b) DP (0.136)



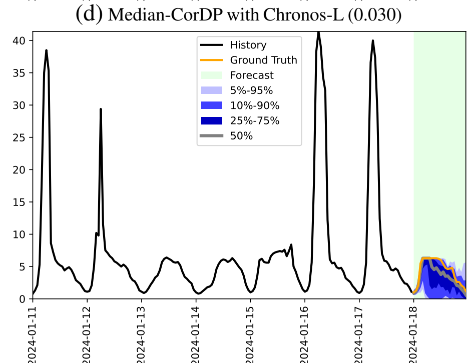
(c) Median-CorDP with Lag-Llama (0.238)



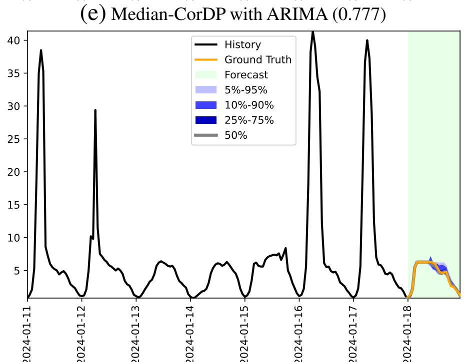
(d) Median-CorDP with Chronos-L (0.030)



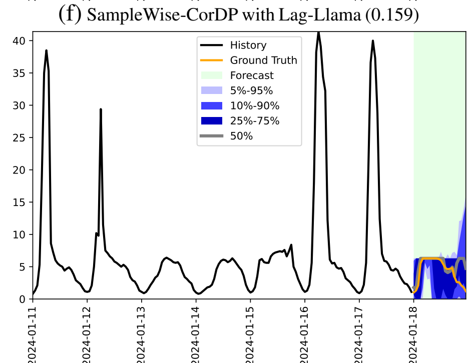
(e) Median-CorDP with ARIMA (0.777)



(f) SampleWise-CorDP with Lag-Llama (0.159)

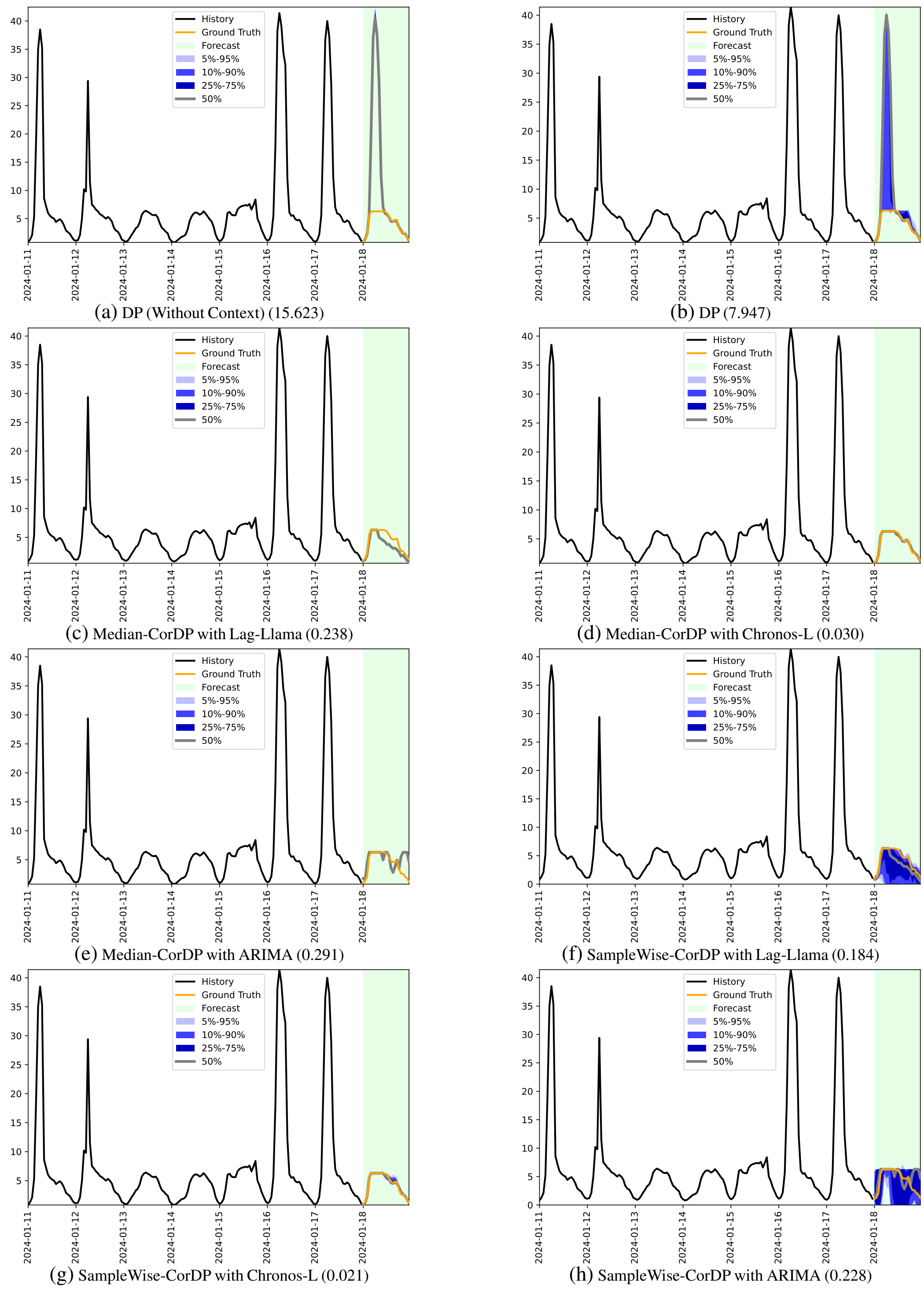


(g) SampleWise-CorDP with Chronos-L (0.027)



(h) SampleWise-CorDP with ARIMA (0.346)

Figure 60: Forecasts of model Qwen2.5-14B-Inst on task *BoundedPredConstraintsBasedOnPredQuantilesTask* (with RCRPS in brackets)



(a) DP (Without Context) (15.623)
(b) DP (7.947)
(c) Median-CorDP with Lag-Llama (0.238)
(d) Median-CorDP with Chronos-L (0.030)
(e) Median-CorDP with ARIMA (0.291)
(f) SampleWise-CorDP with Lag-Llama (0.184)
(g) SampleWise-CorDP with Chronos-L (0.021)
(h) SampleWise-CorDP with ARIMA (0.228)

Figure 61: Forecasts of model Qwen2.5-32B-Inst on task *BoundedPredConstraintsBasedOnPredQuantilesTask* (with RCRPS in brackets)

F Additional Details on IC-DP

F.1 IC-DP Prompt

We use the following prompt for the IC-DP method, where **{example_task_instance.background}**, **{example_task_instance.constraints}**, and **{example_task_instance.scenario}** are replaced by the background, constraints and scenario portions of the context of the example task respectively. **{exam-**

ple_task_history}, **{example_pred_time}** and **{example_task_future}** are replaced by the history, prediction timestamps and the ground truth future values of the examples tasks respectively. **{history}** is replaced by the respective numerical history for the task instance in the format (timestamp, value), and **{context}** is replaced by the respective textual context for the task instance, and **((pred_time))** is replaced with the prediction timesteps. Although this prompt is specialized to the CiK benchmark where contexts are made up of background, constraints and scenario parts, the prompt can be generalized to use any kind of text context.

```
I have a context-aided time series forecasting task for you, where you will be
     given the history of a time series and additional context information, and
     prediction timesteps for which a forecast is required. You are expected to
     factor in any background knowledge,
satisfy any constraints, and respect any scenarios given in the context, and
     output the forecast.
in (timestamp, value) format in between <forecast> and </forecast> tags. You are
     to not include any other information (e.g., comments) in the forecast.

Here is the prompt for an example task:

Here is the context:
<context>\nBackground: {example_task_instance.background}\nConstraints:
     {example_task_instance.constraints}\nScenario:
     {example_task_instance.scenario}\n\n</context>\n\nHere is a historical time
     series in (timestamp, value)
     format:\n<history>{example_task_history}</history>\n\nNow please predict
     the value at the following timestamps: {example_pred_time}.\n

The expected output would be:
<forecast>{example_task_future}</forecast>

Note how the context was incorporated in the forecast. You are expected to do the
     same.
Here is the problem for which you need to return a forecast:

Here is some context about the task.
<context>
{context}
</context>

Here is a historical time series in (timestamp, value) format:
<history>
{history}
</history>

Now please predict the value at the following timestamps: {pred_time}.

Return the forecast in (timestamp, value) format in between <forecast> and
     </forecast> tags.
Do not include any other information (e.g., comments) in the forecast.
```

F.2 Aggregate Results

Table 10 displays the aggregate results of models on CiK, comparing IC-DP and DP.

F.3 Results on various kinds of tasks

Table 11 displays the results of models on various kinds of tasks in CiK, comparing IC-DP and DP.

F.4 Example Forecasts

Model	DP	IC-DP
Llama3.2-1B-Inst	0.396 ± 0.027	$\mathbf{0.337 \pm 0.009}$
Llama3.2-3B-Inst	0.687 ± 0.025	$\mathbf{0.476 \pm 0.018}$
Qwen2.5-0.5B-Inst	0.592 ± 0.027	$\mathbf{0.305 \pm 0.006}$
Qwen2.5-1.5B-Inst	0.616 ± 0.018	$\mathbf{0.273 \pm 0.008}$
Qwen2.5-3B-Inst	0.424 ± 0.017	$\mathbf{0.298 \pm 0.011}$
Qwen2.5-7B-Inst	0.401 ± 0.006	$\mathbf{0.264 \pm 0.012}$
Qwen2.5-14B-Inst	$\mathbf{0.247 \pm 0.006}$	0.270 ± 0.005
Qwen2.5-32B-Inst	0.397 ± 0.008	$\mathbf{0.245 \pm 0.027}$
Qwen2.5-72B-Inst	0.202 ± 0.009	$\mathbf{0.180 \pm 0.014}$
Llama3.3-70B-Inst	0.230 ± 0.006	$\mathbf{0.168 \pm 0.006}$
Llama3.1-405B-Inst	0.173 ± 0.003	$\mathbf{0.129 \pm 0.004}$

Table 10: Results of models with IC-DP on CiK. The best-performing method with each model is in **bold**.

	RoI RCRPS		non-RoI RCRPS		RCRPS of tasks with full RoI		Constraints RCRPS	
Model	DP	IC-DP	DP	IC-DP	DP	IC-DP	DP	IC-DP
Llama3.2-1B-Inst	0.336 ± 0.026	$\mathbf{0.218 \pm 0.006}$	0.248 ± 0.026	$\mathbf{0.187 \pm 0.006}$	0.467 ± 0.041	$\mathbf{0.428 \pm 0.015}$	0.275 ± 0.092	$\mathbf{0.007 \pm 0.031}$
Llama3.2-3B-Inst	0.281 ± 0.013	$\mathbf{0.147 \pm 0.005}$	$\mathbf{0.162 \pm 0.013}$	0.209 ± 0.005	1.004 ± 0.040	$\mathbf{0.679 \pm 0.031}$	1.030 ± 0.090	$\mathbf{0.163 \pm 0.068}$
Llama3.3-70B-Inst	$\mathbf{0.105 \pm 0.003}$	0.134 ± 0.003	0.182 ± 0.003	$\mathbf{0.122 \pm 0.003}$	0.289 ± 0.011	$\mathbf{0.194 \pm 0.010}$	$\mathbf{0.000 \pm 0.024}$	0.025 ± 0.020
Llama3.1-405B-Inst	0.126 ± 0.004	$\mathbf{0.094 \pm 0.004}$	0.150 ± 0.004	$\mathbf{0.115 \pm 0.004}$	0.196 ± 0.005	$\mathbf{0.146 \pm 0.006}$	0.004 ± 0.009	$\mathbf{0.000 \pm 0.012}$
Qwen2.5-0.5B-Inst	0.339 ± 0.010	$\mathbf{0.288 \pm 0.004}$	$\mathbf{0.129 \pm 0.010}$	0.209 ± 0.004	0.836 ± 0.046	$\mathbf{0.343 \pm 0.010}$	0.243 ± 0.103	$\mathbf{0.005 \pm 0.020}$
Qwen2.5-1.5B-Inst	0.317 ± 0.020	$\mathbf{0.224 \pm 0.009}$	0.224 ± 0.020	$\mathbf{0.163 \pm 0.009}$	0.851 ± 0.026	$\mathbf{0.327 \pm 0.011}$	0.706 ± 0.147	$\mathbf{0.023 \pm 0.023}$
Qwen2.5-3B-Inst	0.269 ± 0.015	$\mathbf{0.265 \pm 0.009}$	0.186 ± 0.015	$\mathbf{0.180 \pm 0.009}$	0.558 ± 0.027	$\mathbf{0.349 \pm 0.017}$	0.234 ± 0.056	$\mathbf{0.031 \pm 0.039}$
Qwen2.5-7B-Inst	0.285 ± 0.006	$\mathbf{0.164 \pm 0.007}$	$\mathbf{0.164 \pm 0.006}$	0.187 ± 0.007	0.521 ± 0.009	$\mathbf{0.325 \pm 0.020}$	0.470 ± 0.078	$\mathbf{0.063 \pm 0.045}$
Qwen2.5-14B-Inst	0.162 ± 0.005	$\mathbf{0.099 \pm 0.003}$	$\mathbf{0.146 \pm 0.005}$	0.148 ± 0.003	$\mathbf{0.310 \pm 0.010}$	0.369 ± 0.008	$\mathbf{0.039 \pm 0.015}$	0.455 ± 0.009
Qwen2.5-32B-Inst	$\mathbf{0.116 \pm 0.001}$	0.129 ± 0.003	0.140 ± 0.001	$\mathbf{0.133 \pm 0.003}$	0.580 ± 0.013	$\mathbf{0.323 \pm 0.045}$	0.479 ± 0.019	$\mathbf{0.186 \pm 0.103}$
Qwen2.5-72B-Inst	$\mathbf{0.115 \pm 0.004}$	0.125 ± 0.003	0.138 ± 0.004	$\mathbf{0.113 \pm 0.003}$	0.253 ± 0.015	$\mathbf{0.221 \pm 0.023}$	$\mathbf{0.032 \pm 0.028}$	0.068 ± 0.052

Table 11: Results of models with IC-DP in various groups of tasks in CiK. The best-performing method with each model in every group is in **bold**.

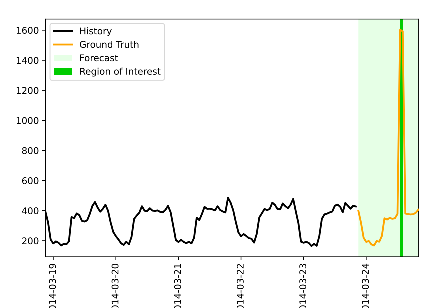
F.4.1 Task: *ElectricityIncreaseInPredictionWithDistractorWithDates*

Background: This is the electricity consumption recorded in Kilowatt (kW) in city A.
Constraints: None.
Scenario: There was a festival in neighbouring cities B and C that resulted in 10 times the usual electricity being consumed there from 2013-05-28 12:00:00 for 2 hours. But this did not affect electricity consumption in city A. Suppose that there is a heat wave in city A from 2013-05-28 12:00:00 for 2 hours, leading to excessive use of air conditioning, and 4 times the usual electricity being consumed.

Figure 62: Context

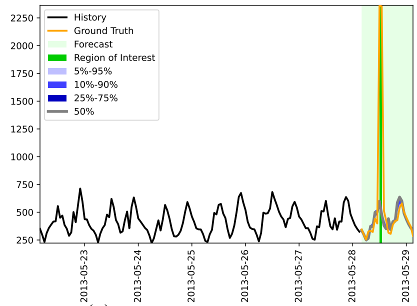
Background: This is the electricity consumption recorded in Kilowatt (kW) in city A.
Constraints: None
Scenario: A brief technical issue in the electricity grid in a nearby city caused a major dip of 75% from 2014-03-24 13:00:00 for 2 hours. This issue has affected many nearby cities, but not this city.Suppose that there is a heat wave in city A from 2014-03-24 13:00:00 for 2 hours, leading to excessive use of air conditioning, and 4 times the usual electricity being consumed.

(a) Context of the In-Context Example Task used with IC-DP experiments

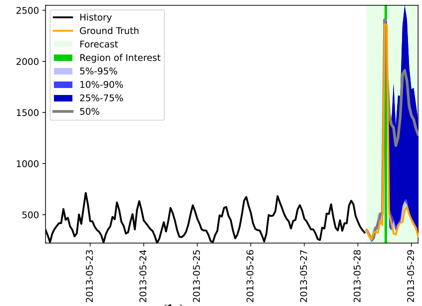


(b) Historical and Future Data of the In-Context Example Task used with IC-DP experiments

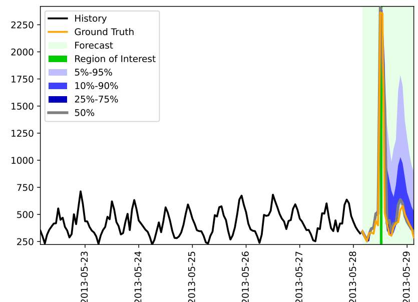
Figure 63: In-Context Example Task used with IC-DP experiments



(a) DP (Without Context) (0.037)

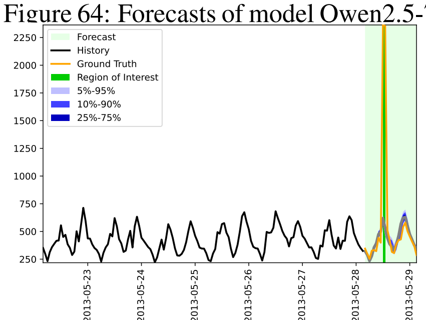


(b) DP (0.010)

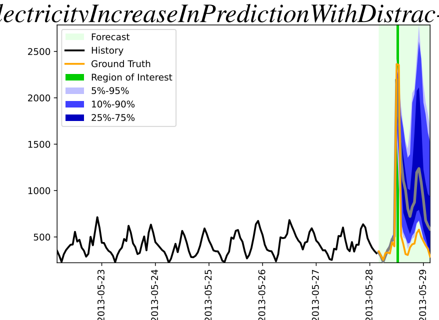


(c) IC-DP (0.003)

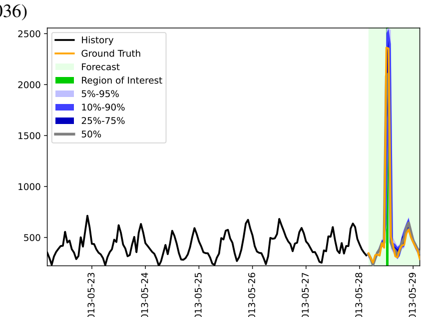
Figure 64: Forecasts of model Qwen2.5-72B-Inst on task *ElectricityIncreaseInPredictionWithDistrac-*



(a) DP (Without Context) (0.036)



(b) DP (0.010)



(c) IC-DP (0.004)

Figure 65: Forecasts of model Llama3.1-405B-Inst on task *ElectricityIncreaseInPredictionWithDistractorWithDates* (with RCRPS in brackets)

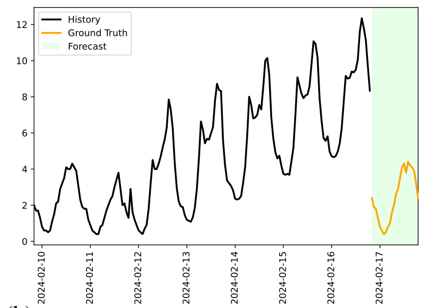
F.4.2 Task: *SensorTrendAccumulationTask*

> Background: This series represents the occupancy rate (%) captured by a highway sensor. The sensor had a calibration problem starting from 2024-01-11 12:00:00 which resulted in an additive trend in the series that increases by 0.0874 at every hour. At timestep 2024-01-18 00:00:00, the sensor was repaired and this additive trend will disappear.
> Constraints: None
> Scenario: None

Figure 66: Context

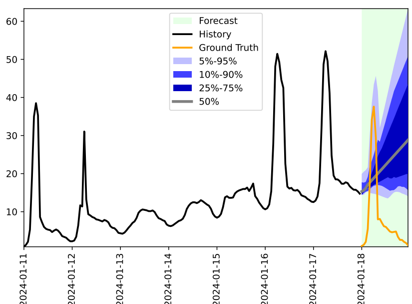
Background: This series represents the occupancy rate (%) captured by a highway sensor. The sensor had a calibration problem starting from 2024-02-12 13:00:00 which resulted in an additive trend in the series that increases by 0.0489 at every hour. At timestep 2024-02-16 20:00:00, the sensor was repaired and this additive trend will disappear.
Constraints: None
Scenario: None

(a) Context of the In-Context Example Task used with IC-DP experiments

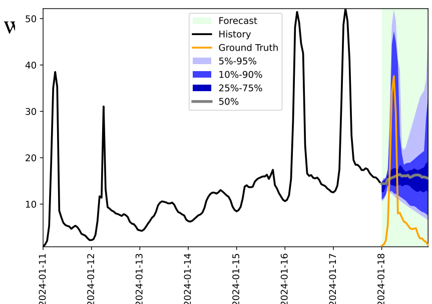


(b) Historical and Future Data of the In-Context Example Task used with IC-DP experiments

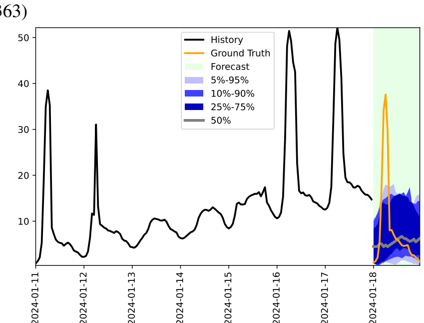
Example Task used w



(a) DP (Without Context) (1.363)

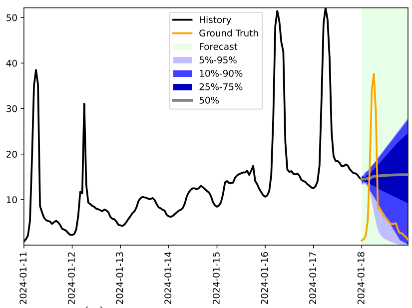


(b) DP (0.856)

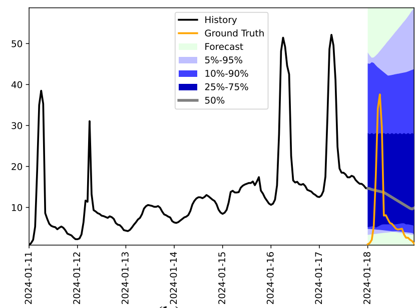


(c) IC-DP (0.455)

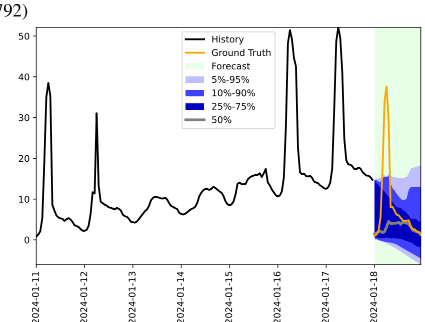
Figure 68: Forecasts of model Llama3-8B-Inst on task *SensorTrendAccumulationTask* (with RCRPS in brackets)



(a) DP (Without Context) (0.792)

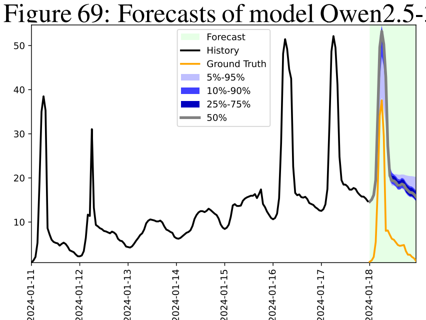


(b) DP (0.616)

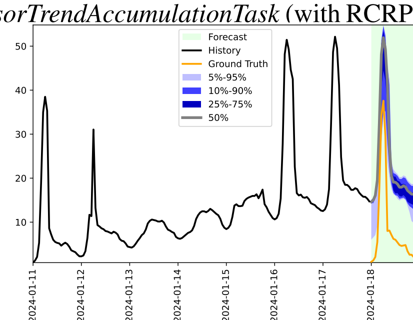


(c) IC-DP (0.427)

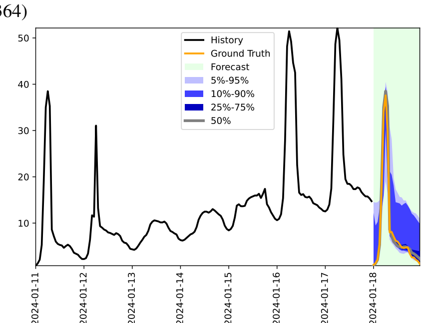
Figure 69: Forecasts of model Qwen2.5-3B-Inst on task *SensorTrendAccumulationTask* (with RCRPS



(a) DP (Without Context) (1.364)



(b) DP (1.144)



(c) IC-DP (0.073)

Figure 70: Forecasts of model Llama3.1-405B-Inst on task *SensorTrendAccumulationTask* (with RCRPS in brackets)

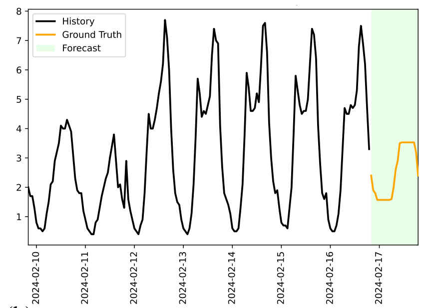
F.4.3 Task: *BoundedPredConstraintsBasedOnPredQuantilesTask*

Background: None Constraints: Suppose that in the forecast, the values are bounded above by 6.29. Scenario: None

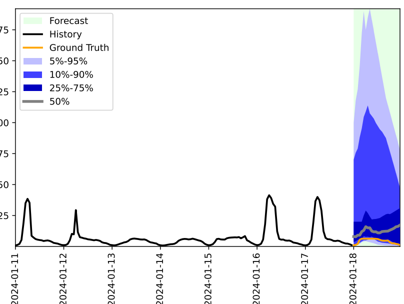
Figure 71: Context

Background: None
Constraints: Suppose that in the forecast, the values are bounded below by 1.57, the values are bounded above by 3.53.
Scenario: None

(a) Context of the In-Context Example Task used with IC-DP experiments

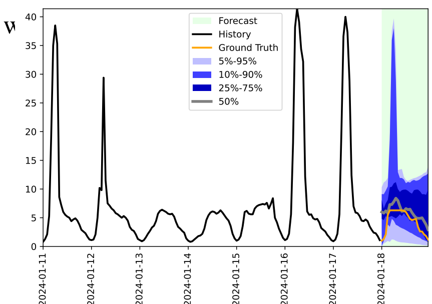


(b) Historical and Future Data of the In-Context Example Task used with IC-DP experiments

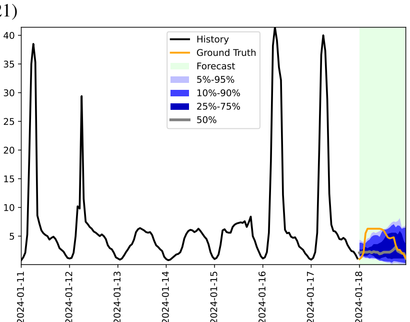


(a) DP (Without Context) (15.021)

Example Task used w

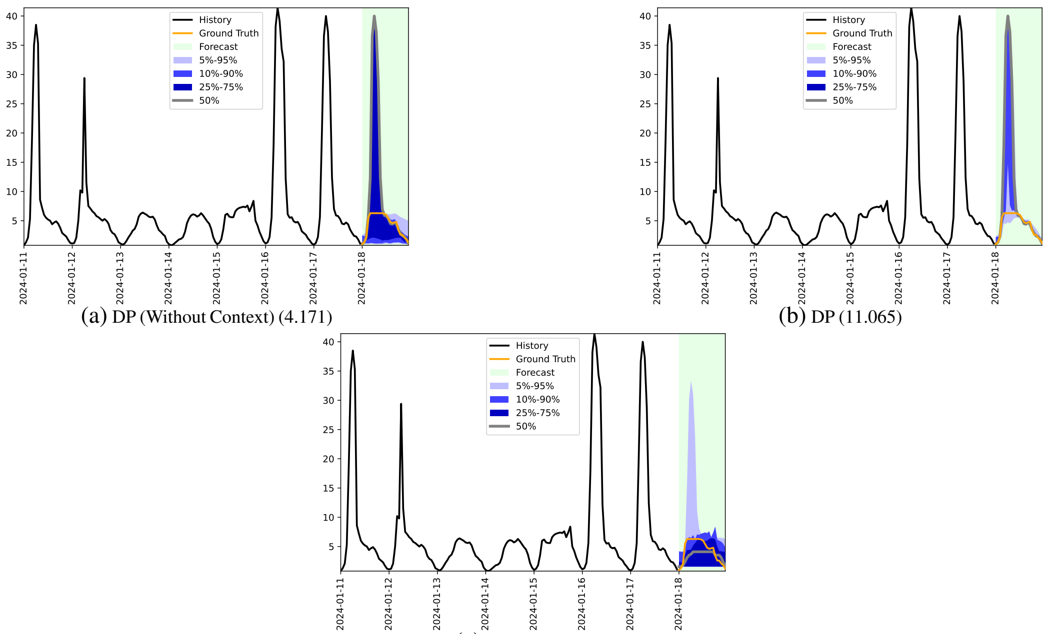


(b) DP (2.784)



(c) IC-DP (0.405)

Figure 73: Forecasts of model Llama3.2-1B-Inst on task *BoundedPredConstraintsBasedOnPredQuantilesTask* (with RCRPS in brackets)

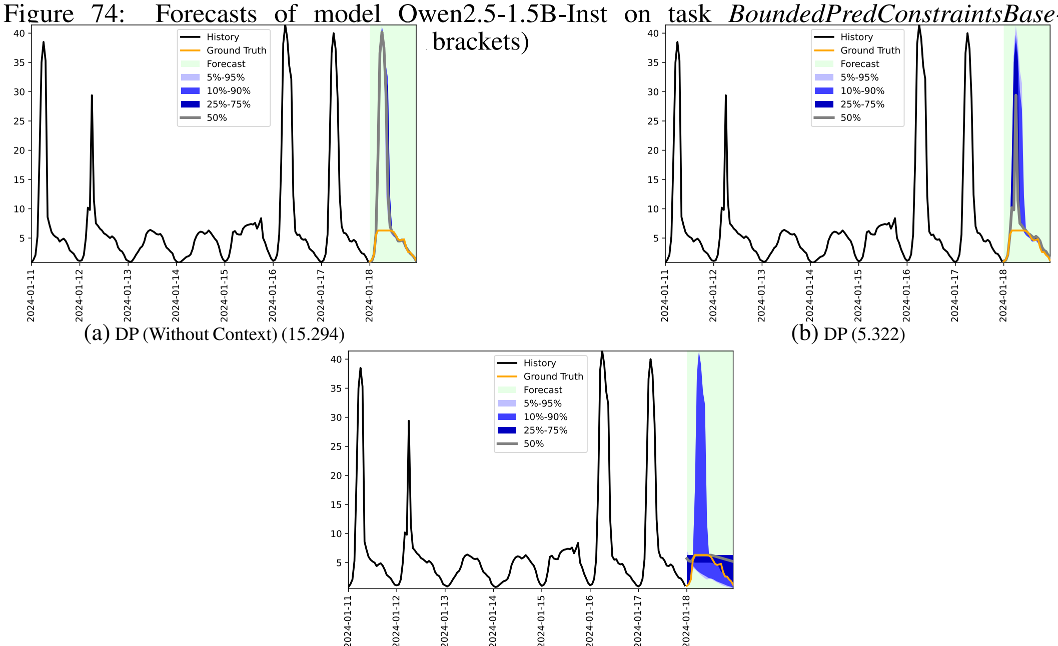


(a) DP (Without Context) (4.171)

(b) DP (11.065)

(c) IC-DP (0.447)

Figure 74: Forecasts of model Qwen2.5-1.5B-Inst on task *BoundedPredConstraintsBase-* brackets)



(a) DP (Without Context) (15.294)

(b) DP (5.322)

(c) IC-DP (0.771)

Figure 75: Forecasts of model Qwen2.5-7B-Inst on task *BoundedPredConstraintsBasedOnPredQuantilesTask* (with RCRPS in brackets)

G Additional Details on RouteDP

G.1 RouteDP Prompt

To predict the difficulty of a task, we use the below prompt, where **{direct_prompt}** is replaced by the instantiated Direct Prompt (DP) prompt, which contains the context, historical time series and prediction timesteps of the task, as used in Williams et al. [48].

```
{direct_prompt}
You are given a forecasting task with full contextual information.
Please rate the task as easy or hard.
Difficulty:
```

Given all 71 tasks from the CiK benchmark, we first run the designated Router LLM to predict the difficulty of a task. In particular, we use constrained decoding to limit the outputs to either "easy" or "hard".

Then, to route tasks, given a k number of tasks to send to the large model, we dispatch the top-k tasks considered hardest according to $P(hard)$ to the larger LLM, and dispatch the rest to the main model.

G.2 Extended Results

Main Model	Router	Percentage of tasks sent to large model					
		0%	20%	40%	60%	80%	100%
Qwen2.5-0.5B-Inst							
	Qwen2.5-0.5B-Inst	0.592 ± 0.027	0.316 ± 0.027	0.222 ± 0.005	0.206 ± 0.005	0.199 ± 0.004	0.173 ± 0.003
	Qwen2.5-1.5B-Inst	0.592 ± 0.027	0.504 ± 0.009	0.449 ± 0.007	0.404 ± 0.004	0.407 ± 0.004	0.173 ± 0.003
	Qwen2.5-3B-Inst	0.592 ± 0.027	0.507 ± 0.026	0.490 ± 0.026	0.393 ± 0.003	0.282 ± 0.003	0.173 ± 0.003
	Qwen2.5-7B-Inst	0.592 ± 0.027	0.510 ± 0.010	0.437 ± 0.007	0.412 ± 0.004	0.181 ± 0.004	0.173 ± 0.003
	Qwen2.5-14B-Inst	0.592 ± 0.027	0.581 ± 0.027	0.439 ± 0.027	0.324 ± 0.027	0.187 ± 0.004	0.173 ± 0.003
	Qwen2.5-32B-Inst	0.592 ± 0.027	0.383 ± 0.010	0.368 ± 0.008	0.230 ± 0.006	0.196 ± 0.004	0.173 ± 0.003
	Qwen2.5-72B-Inst	0.592 ± 0.027	0.509 ± 0.010	0.395 ± 0.009	0.287 ± 0.009	0.243 ± 0.009	0.173 ± 0.003
Qwen2.5-1.5B-Inst							
	Qwen2.5-0.5B-Inst	0.616 ± 0.018	0.436 ± 0.016	0.258 ± 0.005	0.231 ± 0.005	0.210 ± 0.004	0.173 ± 0.003
	Qwen2.5-1.5B-Inst	0.616 ± 0.018	0.466 ± 0.016	0.300 ± 0.016	0.212 ± 0.005	0.210 ± 0.005	0.173 ± 0.003
	Qwen2.5-3B-Inst	0.616 ± 0.018	0.375 ± 0.009	0.349 ± 0.009	0.196 ± 0.004	0.181 ± 0.004	0.173 ± 0.003
	Qwen2.5-7B-Inst	0.616 ± 0.018	0.481 ± 0.018	0.288 ± 0.016	0.235 ± 0.005	0.188 ± 0.004	0.173 ± 0.003
	Qwen2.5-14B-Inst	0.616 ± 0.018	0.598 ± 0.017	0.536 ± 0.016	0.523 ± 0.015	0.214 ± 0.004	0.173 ± 0.003
	Qwen2.5-32B-Inst	0.616 ± 0.018	0.441 ± 0.017	0.356 ± 0.017	0.256 ± 0.009	0.210 ± 0.004	0.173 ± 0.003
	Qwen2.5-72B-Inst	0.616 ± 0.018	0.484 ± 0.017	0.448 ± 0.017	0.445 ± 0.017	0.375 ± 0.015	0.173 ± 0.003
Qwen2.5-3B-Inst							
	Qwen2.5-0.5B-Inst	0.424 ± 0.017	0.338 ± 0.014	0.301 ± 0.011	0.281 ± 0.011	0.208 ± 0.004	0.173 ± 0.003
	Qwen2.5-1.5B-Inst	0.424 ± 0.017	0.383 ± 0.014	0.309 ± 0.012	0.260 ± 0.012	0.262 ± 0.012	0.173 ± 0.003
	Qwen2.5-3B-Inst	0.424 ± 0.017	0.315 ± 0.015	0.254 ± 0.011	0.203 ± 0.006	0.188 ± 0.005	0.173 ± 0.003
	Qwen2.5-7B-Inst	0.424 ± 0.017	0.382 ± 0.015	0.285 ± 0.012	0.276 ± 0.012	0.228 ± 0.010	0.173 ± 0.003
	Qwen2.5-14B-Inst	0.424 ± 0.017	0.402 ± 0.017	0.340 ± 0.015	0.329 ± 0.015	0.246 ± 0.011	0.173 ± 0.003
	Qwen2.5-32B-Inst	0.424 ± 0.017	0.359 ± 0.014	0.326 ± 0.013	0.295 ± 0.012	0.262 ± 0.011	0.173 ± 0.003
	Qwen2.5-72B-Inst	0.424 ± 0.017	0.392 ± 0.015	0.345 ± 0.015	0.338 ± 0.014	0.289 ± 0.013	0.173 ± 0.003
Qwen2.5-7B-Inst							
	Qwen2.5-0.5B-Inst	0.401 ± 0.006	0.364 ± 0.005	0.238 ± 0.004	0.229 ± 0.004	0.208 ± 0.004	0.173 ± 0.003
	Qwen2.5-1.5B-Inst	0.401 ± 0.006	0.263 ± 0.006	0.222 ± 0.005	0.183 ± 0.004	0.181 ± 0.004	0.173 ± 0.003
	Qwen2.5-3B-Inst	0.401 ± 0.006	0.338 ± 0.004	0.314 ± 0.004	0.179 ± 0.004	0.174 ± 0.004	0.173 ± 0.003
	Qwen2.5-7B-Inst	0.401 ± 0.006	0.260 ± 0.006	0.199 ± 0.005	0.191 ± 0.004	0.188 ± 0.004	0.173 ± 0.003
	Qwen2.5-14B-Inst	0.401 ± 0.006	0.384 ± 0.006	0.351 ± 0.006	0.343 ± 0.005	0.194 ± 0.004	0.173 ± 0.003
	Qwen2.5-32B-Inst	0.401 ± 0.006	0.260 ± 0.006	0.240 ± 0.005	0.231 ± 0.004	0.206 ± 0.004	0.173 ± 0.003
	Qwen2.5-72B-Inst	0.401 ± 0.006	0.267 ± 0.006	0.246 ± 0.006	0.244 ± 0.006	0.214 ± 0.005	0.173 ± 0.003
Qwen2.5-14B-Inst							
	Qwen2.5-0.5B-Inst	0.247 ± 0.006	0.208 ± 0.004	0.202 ± 0.004	0.199 ± 0.004	0.194 ± 0.004	0.173 ± 0.003
	Qwen2.5-1.5B-Inst	0.247 ± 0.006	0.246 ± 0.006	0.220 ± 0.006	0.196 ± 0.006	0.199 ± 0.006	0.173 ± 0.003
	Qwen2.5-3B-Inst	0.247 ± 0.006	0.206 ± 0.006	0.204 ± 0.006	0.192 ± 0.006	0.189 ± 0.004	0.173 ± 0.003
	Qwen2.5-7B-Inst	0.247 ± 0.006	0.234 ± 0.007	0.197 ± 0.006	0.198 ± 0.006	0.179 ± 0.003	0.173 ± 0.003
	Qwen2.5-14B-Inst	0.247 ± 0.006	0.237 ± 0.006	0.203 ± 0.006	0.200 ± 0.004	0.176 ± 0.003	0.173 ± 0.003
	Qwen2.5-32B-Inst	0.247 ± 0.006	0.230 ± 0.005	0.220 ± 0.005	0.195 ± 0.003	0.193 ± 0.003	0.173 ± 0.003
	Qwen2.5-72B-Inst	0.247 ± 0.006	0.238 ± 0.006	0.218 ± 0.005	0.203 ± 0.004	0.202 ± 0.004	0.173 ± 0.003
Qwen2.5-32B-Inst							
	Qwen2.5-0.5B-Inst	0.397 ± 0.008	0.296 ± 0.005	0.171 ± 0.003	0.172 ± 0.004	0.172 ± 0.003	0.173 ± 0.003
	Qwen2.5-1.5B-Inst	0.397 ± 0.008	0.278 ± 0.008	0.272 ± 0.008	0.264 ± 0.007	0.266 ± 0.007	0.173 ± 0.003
	Qwen2.5-3B-Inst	0.397 ± 0.008	0.390 ± 0.007	0.384 ± 0.007	0.265 ± 0.007	0.218 ± 0.006	0.173 ± 0.003
	Qwen2.5-7B-Inst	0.397 ± 0.008	0.276 ± 0.008	0.273 ± 0.008	0.265 ± 0.007	0.175 ± 0.003	0.173 ± 0.003
	Qwen2.5-14B-Inst	0.397 ± 0.008	0.397 ± 0.008	0.361 ± 0.007	0.310 ± 0.005	0.177 ± 0.003	0.173 ± 0.003
	Qwen2.5-32B-Inst	0.397 ± 0.008	0.240 ± 0.007	0.237 ± 0.007	0.185 ± 0.003	0.181 ± 0.004	0.173 ± 0.003
	Qwen2.5-72B-Inst	0.397 ± 0.008	0.284 ± 0.008	0.236 ± 0.007	0.191 ± 0.005	0.193 ± 0.006	0.173 ± 0.003
Qwen2.5-72B-Inst							
	Qwen2.5-0.5B-Inst	0.202 ± 0.009	0.167 ± 0.007	0.156 ± 0.004	0.158 ± 0.004	0.165 ± 0.004	0.173 ± 0.003
	Qwen2.5-1.5B-Inst	0.202 ± 0.009	0.184 ± 0.006	0.181 ± 0.006	0.185 ± 0.006	0.194 ± 0.006	0.173 ± 0.003
	Qwen2.5-3B-Inst	0.202 ± 0.009	0.207 ± 0.009	0.210 ± 0.009	0.189 ± 0.006	0.178 ± 0.004	0.173 ± 0.003
	Qwen2.5-7B-Inst	0.202 ± 0.009	0.187 ± 0.006	0.185 ± 0.006	0.192 ± 0.006	0.175 ± 0.003	0.173 ± 0.003
	Qwen2.5-14B-Inst	0.202 ± 0.009	0.207 ± 0.009	0.202 ± 0.009	0.190 ± 0.008	0.175 ± 0.003	0.173 ± 0.003
	Qwen2.5-32B-Inst	0.202 ± 0.009	0.183 ± 0.005	0.186 ± 0.005	0.180 ± 0.004	0.175 ± 0.004	0.173 ± 0.003
	Qwen2.5-72B-Inst	0.202 ± 0.009	0.198 ± 0.006	0.188 ± 0.005	0.185 ± 0.004	0.180 ± 0.004	0.173 ± 0.003

Table 12: Average RCRPS of main models routed with different routers, as the percentage of tasks sent to the large model increases. The means are accompanied by standard errors.

Router	Qwen2.5-0.5B-Inst	Qwen2.5 1.5B	Qwen2.5 3B	Qwen2.5 7B	Qwen2.5 14B	Qwen2.5 32B	Qwen2.5-72B-Inst
Qwen2.5-0.5B-Inst	66.76	48.83	13.50	29.05	22.03	68.59	67.59
Qwen2.5-1.5B-Inst	1.40	40.53	4.67	55.13	2.63	19.65	12.61
Qwen2.5-3B-Inst	3.10	46.41	45.41	23.05	22.23	3.47	1.23
Qwen2.5-7B-Inst	7.95	35.33	7.41	55.72	15.73	26.63	6.71
Qwen2.5-14B-Inst	10.12	4.62	0.00	5.11	12.14	7.66	3.77
Qwen2.5-32B-Inst	36.96	39.04	6.52	51.15	7.79	58.45	14.86
Qwen2.5-72B-Inst	9.73	3.35	0.00	29.19	0.49	34.67	3.34

Table 13: Area captured by each router for each main model, between the main model's own random and ideal routing curves. Values of different routers across the same main model are comparable.

G.3 Plots showcasing the area captured by different router models

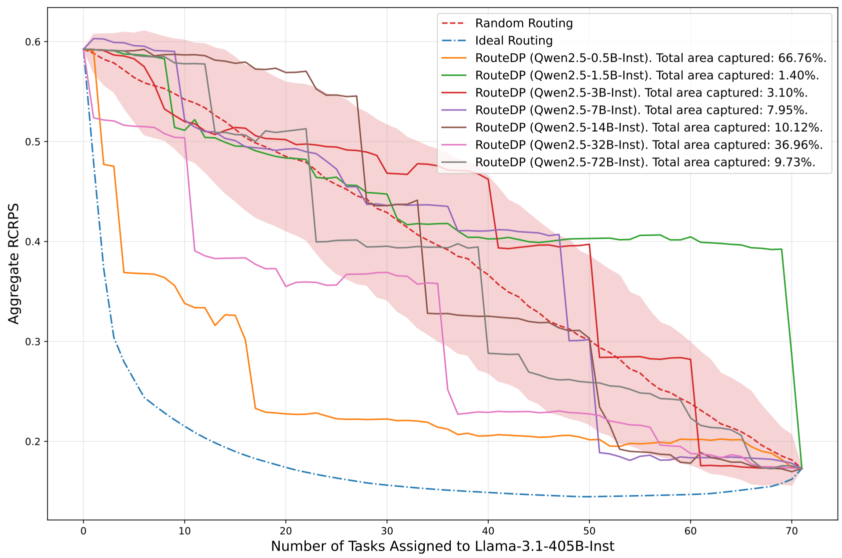


Figure 76: Random, ideal and router curves with Qwen2.5-0.5B-Inst as the main model. The shaded region represents the distribution of 100 random assignment trajectories.

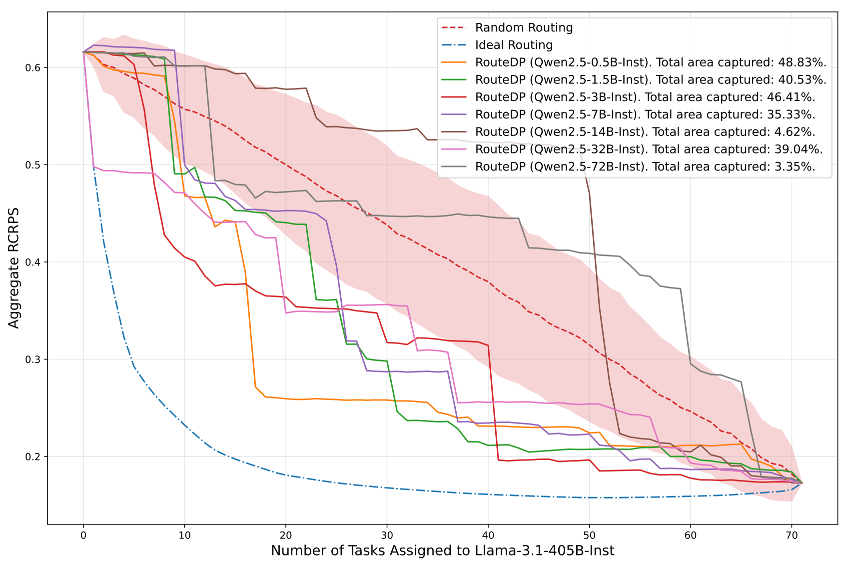


Figure 77: Random, ideal and router curves with Qwen2.5-1.5B-Inst as the main model. The shaded region represents the distribution of 100 random assignment trajectories.

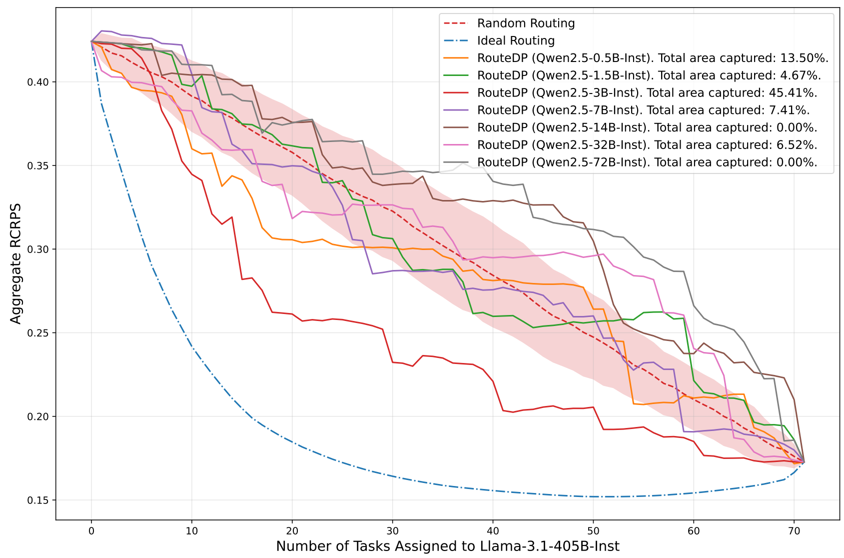


Figure 78: Random, ideal and router curves with Qwen2.5-3B-Inst as the main model. The shaded region represents the distribution of 100 random assignment trajectories.

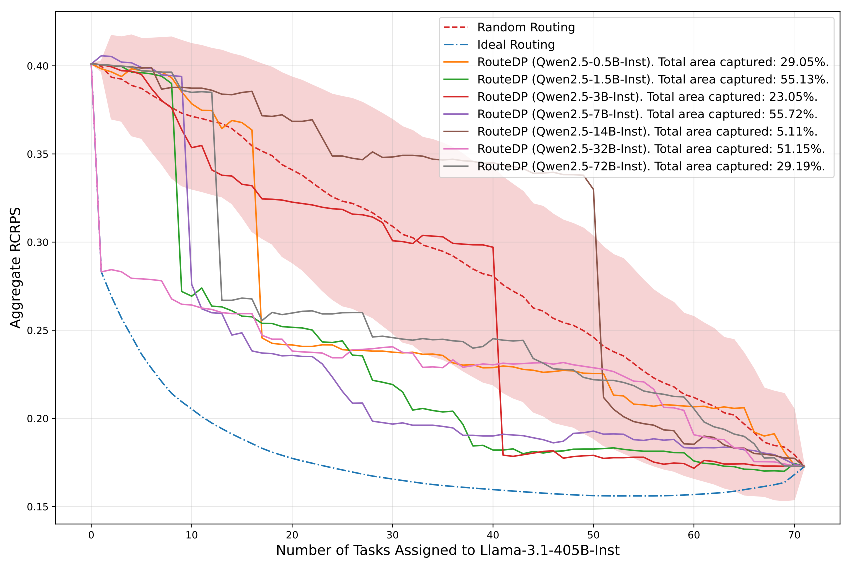


Figure 79: Random, ideal and router curves with Qwen2.5-7B-Inst as the main model. The shaded region represents the distribution of 100 random assignment trajectories.

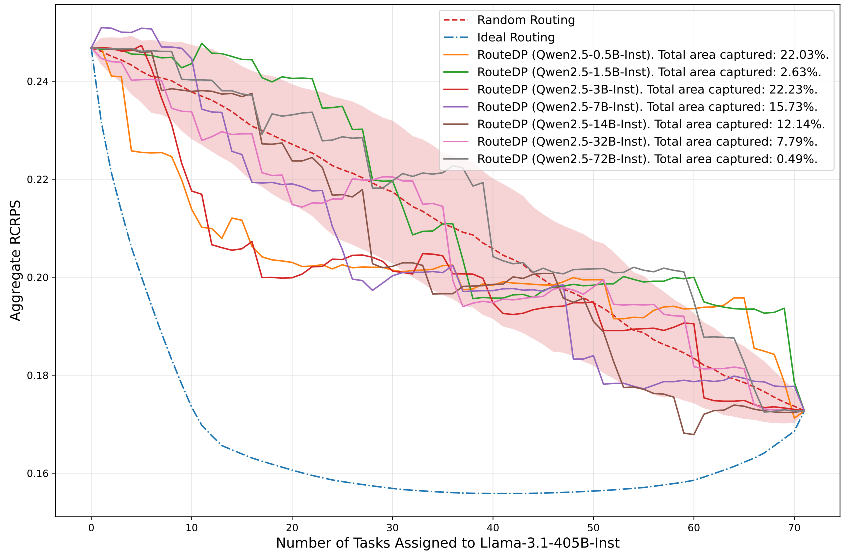


Figure 80: Random, ideal and router curves with Qwen2.5-14B-Inst as the main model. The shaded region represents the distribution of 100 random assignment trajectories.

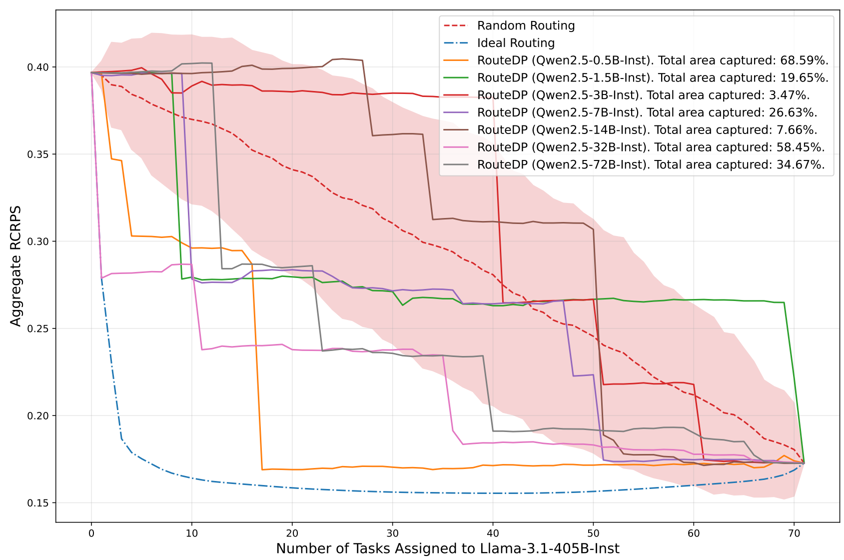


Figure 81: Random, ideal and router curves with Qwen2.5-32B-Inst as the main model. The shaded region represents the distribution of 100 random assignment trajectories.

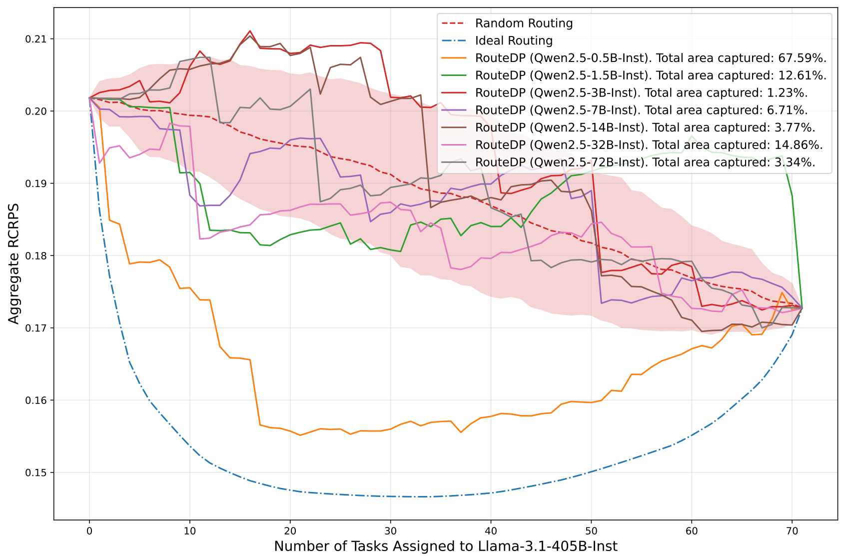


Figure 82: Random, ideal and router curves with Qwen2.5-72B-Inst as the main model. The shaded region represents the distribution of 100 random assignment trajectories.

H Implementation Details

To evaluate our models on the CiK benchmark, we use the official codebase of CiK at `https://github.com/ServiceNow/context-is-key-forecasting`. We use the same codebase to run model on the Direct Prompt (DP) method and the quantitative baselines benchmarked for CorDP. For completeness, we provide the details here. Code for all proposed methods will be released on acceptance, with instructions to reproduce all experiments.

H.1 LLMs

We self-host the other models with the respective official HuggingFace models: Llama3.2-1B-Inst (`https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct`), Llama-3.2-3B-Inst (`https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct`), Llama-3-8B-Inst (`https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`), Qwen2.5-0.5B-Inst (`https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct`), Qwen2.5-1.5B-Inst (`https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct`), Qwen2.5-3B-Inst (`https://huggingface.co/Qwen/Qwen2.5-3B-Instruct`), Qwen2.5-7B-Inst (`https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`), Qwen2.5-14B-Inst (`https://huggingface.co/Qwen/Qwen2.5-14B-Instruct`), Qwen2.5-32B-Inst (`https://huggingface.co/Qwen/Qwen2.5-32B-Instruct`). We use an appropriate number of H100 GPUs for each model. This ranged from 1 GPU (Models below 7B), 2 GPUs (7B, 14B Models) and 4 GPUs (32B Models).

Due to compute restrictions, for all our experiments involving Llama-3.1-405B-Inst, Llama3.3-70B-Inst and Qwen2.5-72B-Inst, we use OpenRouter endpoints at `https://openrouter.ai/meta-llama/llama-3.1-405b-instruct`, `https://openrouter.ai/meta-llama/llama-3.3-70b-instruct` and `https://openrouter.ai/qwen/qwen-2.5-72b-instruct` respectively.

For all the above LLMs, we use the below prompt for the Direct Prompt method, as given in `https://github.com/ServiceNow/context-is-key-forecasting`. **context**, **history** and **pred_time** are replaced by the respective textual context, numerical history and timestamps for which a forecast is required.

```
I have a time series forecasting task for you.

Here is some context about the task. Make sure to factor in any background knowledge,
satisfy any constraints, and respect any scenarios.
<context>
((context))
</context>

Here is a historical time series in (timestamp, value) format:
<history>
((history))
</history>

Now please predict the value at the following timestamps: ((pred_time)).

Return the forecast in (timestamp, value) format in between <forecast> and </forecast> tags.
Do not include any other information (e.g., comments) in the forecast.

Example:
<history>
(t1, v1)
(t2, v2)
(t3, v3)
</history>
<forecast>
(t4, v4)
(t5, v5)
</forecast>
```

H.2 Lag-Llama

We use the publicly available implementation of Lag-Llama [37] following the instructions at `https://github.com/time-series-foundation-models/`, on a single H100 GPU.

H.3 Chronos

We use the publicly available implementation of Chronos-Large [2] following the instructions at at `https://github.com/amazon-science/chronos-forecasting` on a single H100 GPU.

H.4 ARIMA

We used the implementation of ARIMA from the `forecast` R package, using `rpy2`. Results are computed using the `auto.arima` method. We reran the model with restricted parameter and disabled seasonality if the ARIMA fit failed.

I Future Work

Our work opens up several directions of research, highlighting the room for potential improvements in context-aided forecasting methods without any training. To start with, combining the proposed strategies can yield interesting analyses, revealing further interesting insights about the capabilities of LLMs and potentially improving their performance. Next, while we limit our scope to studying these strategies with the DP method, it is also interesting to study them in the context of other methods such as LLMP [38] and LLMTime [17]. Next, exploring the usefulness of these strategies in more unconstrained setups where for e.g. all context may not be relevant, or where the context is long, are also interesting directions, however first requires the development of datasets where the contexts have the respective properties to test for. Our scope is also limited to improving the *zero-shot* performance of LLMs; exploring these strategies in the other paradigm of training-based methods could be useful. In particular, moving to training-based methods can also broaden the scope of these strategies e.g. training the router in RouteDP or training a model to use any base forecaster in CorDP. Finally, while these strategies improve the zero-shot performance of LLMs, the high cost of LLMs compared to canonical forecasting methods still limit the applicability of LLMs in deployment; these strategies and studies must in-turn be used to develop more efficient models from the ground-up, keeping the requirements of the respective forecasting application in mind.