

Supplementary Material for: Kairos: Redefining Event and Time Prediction with Language Modeling

Akash Gupta
Amazon, India

AKASHGGU@AMAZON.COM

Sahil Verma
Amazon, India

VRSAHIL@AMAZON.COM

Sumit Bisht
Amazon, India

SUBISHT@AMAZON.COM

Nitika Verma
Amazon, India

NITIKAV@AMAZON.COM

Purav Aggarwal
Amazon, India

AGGAP@AMAZON.COM

Venkat Phanindra Kumar Grandhi
Amazon, India

PHANINDR@AMAZON.COM

Abhishek Persad
Amazon, India

PERSADAP@AMAZON.COM

Editors: Hung-yi Lee and Tongliang Liu

1. Additional Results

1.1. Impact of scaling number of generations (M) on Long-Term Forecasting

Increasing the number of generations in the self-consistency paradigm has been shown to improve performance (Wang et al., 2022). In this study, we investigate the behavior of Kairos as the number of generations for self-consistency with consensus decoding are increased. Figure 1 demonstrates a similar scaling effect, with the OTD value reducing to 13.79 when using 50 generations. While this improvement comes at the cost of increased inference times, it indicates that the long-term performance of Kairos can be further enhanced beyond the results presented in the main paper .

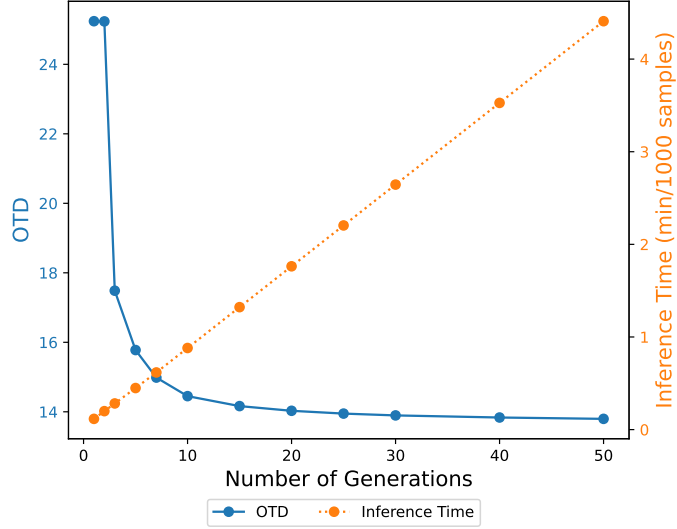


Figure 1: OTD vs Number of generations on SO-400K dataset

Table 1: Impact of B on SO-400K dataset.

B	$E_{Acc}(\%)$	T_{RMSE}
32	32.14	2.429
128	32.18	2.338
512	30.07	2.382
2048	27.76	2.368

1.2. Impact of number of bins (B) on Short-Term Forecasting

To understand the impact of number of bins on Kairos’ performance, we evaluate model performance on SO-400K dataset with different values of B viz. 32, 128, 512 and 2048 (Table 1). Increasing B from 32 to 128 improves T_{RMSE} (2.429 to 2.338) by reducing quantization error, which is theoretically bounded by $|\tau - \hat{\tau}| \leq \frac{\tau_{\max}}{2B}$ where τ is the original time value and $\hat{\tau}$ is the quantized time value. This bound follows from uniform quantization: with B bins of width $\frac{\tau_{\max}}{B}$, the maximum error occurs when a value falls halfway between the quantization points, giving a maximum error of half the bin width. However, increasing the number of bins further to 512 and 2,048 degrades performance due to data sparsity issues (Ansari et al., 2024). This effect extends to E_{acc} , which drops from 32.18% at 128 bins to 27.76% at 2,048 bins, demonstrating that optimal time discretization requires balancing quantization error against sufficient examples per bin for effective learning.

1.3. Impact of number of model parameters on CTES Forecasting

Table 2 presents the change in model performance as number of parameters in the base model increases on the SO-400K dataset. As the model size increases, the E_{Acc} of the model

improves by 11% with improvement of 2% and OTD improvement of 8%. This is because the increased model capacity increases the ability of the model to learn complex patterns and consequently the performance improves for both long-term and short-term forecasting, in line with scaling laws observed in language modeling (Hoffmann et al., 2022).

Table 2: Impact of model parameters on SO-400K dataset.

Layers	Heads	d_model	Parameter Count	$E_{Acc}(\%)$	T_{RMSE}	OTD
2	2	16	29.8K	32.18	2.338	25.24
2	2	32	71.7K	33.36	2.340	24.81
4	4	64	292K	35.67	2.300	23.67
8	8	128	1.8M	36.04	2.290	23.17

1.4. Comparison of Kairos across different time intervals

Table 3 compares THP and Kairos across different time intervals, organizing predictions into quantile bins (from 0-99 in increments of 5) with smaller bins being more frequent than the larger ones. We observed a consistent pattern where prediction performance deteriorates as inter-arrival times increase for both models, which is expected since larger time intervals typically correspond to infrequent events that are harder to predict accurately. However, Kairos demonstrates significantly better resilience to this performance degradation with Kairos outperforming THP with a lower degradation across both time and event prediction performance.

Table 3: Kairos vs THP across different time intervals.

Time Bin Quantile	Kairos T_{RMSE}	THP T_{RMSE}	Kairos E_{acc}	THP E_{acc}
0-5 (smallest)	0.12	0.13	85%	80%
40-50 (medium)	0.45	0.53	70%	55%
95-99 (largest)	1.8	9.62	51%	28%

References

- 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.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.