Nicolas van Schaik (n.j.i.vanschaik@student.tudelft.nl)

TIME'S UP: ROBUST WATERMARKING IN LARGE LANGUAGE MODELS FOR TIME SERIES GENERATION

Introduction

Background:

- LLM usage for time series generation is rapidly increasing [1][2][3] -> need for detection and recognition to prevent harm

Research gap: emerging models so no research on watermarks for time series foundation models

Research question: How do you develop a robust watermarking method for time series foundation models?

Methodology

- Apply several conventional LLM watermarking methods [4] to time series foundation models
- Implement original watermarking algorithm, the Heads Tails Watermark (HTW) and compare its performance to the others
- Performance comparison for three key factors: prediction quality (sMAPE), detection confidence (z-score) and robustness to post-editing attacks (z-score).

Meet the Contenders:

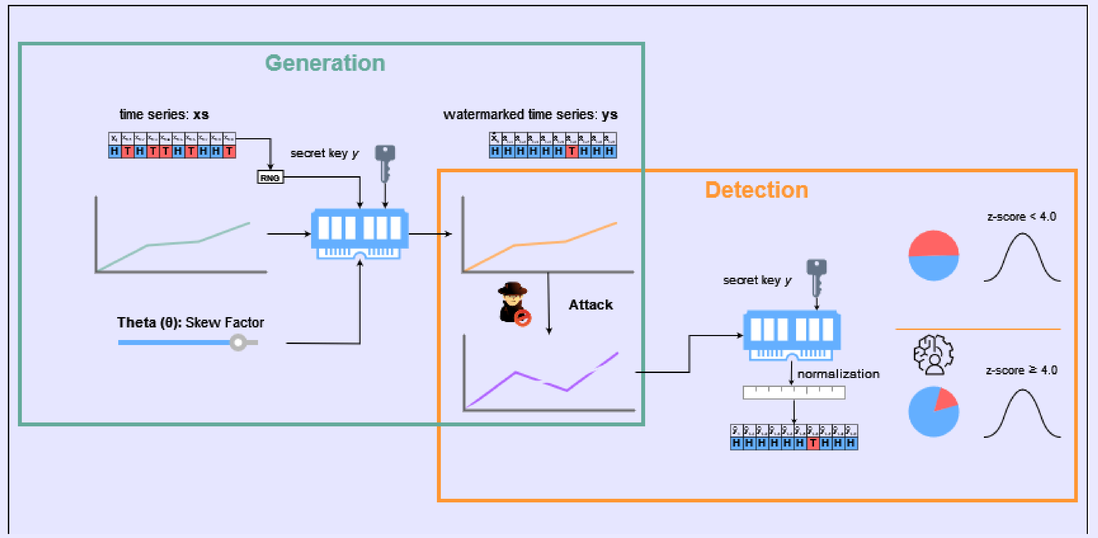
- KGW: most famous LLM watermark with red-green list to alter estimated likelihood
- EXP: watermark implemented by OpenAI's Scott Aaronson that uses secret key and pseudo-random function
- HTW: original implementation that directly embeds signal in numeric structure of the series

Experiments performed with Chronos-base-200M and Lag-Llama

[1] Das, A. Kong, W. Sen, R. Zhou, Y. 2024. A decoder-only foundation model for time-series forecasting https://arxiv.org/abs/2310.10688
[2] Fatir, A. et al. 2024. Chronos: Learning the language of Time Series. https://arxiv.org/pdf/2403.07815
[3] Rasul, K. et al. 2024. Lag-llama: towards foundation models for probabilistic time series forecasting. https://arxiv.org/pdf/2310.08278
[4] Kirchenbauer et al. 2023. A Watermark for Large Language Models. https://arxiv.org/abs/2301.10226

Heads Tails Watermark

The HTW algorithm watermarks a time series by first setting heads and tails targets based on a desired skew factor θ. It then processes each element, normalizing and pseudo-randomly transforming it, based on a secret key y, to append to the output series while adjusting elements to meet the heads or tails target counts.



Results

Prediction Quality

L	12	24	36	48	60	72
HTW	**0.04**	**0.02**	**0.03**	0.02	**0.02**	0.03
EXP	2.65	0.57	0.92	3.06	0.07	**-0.43**
KGW	4.03	4.54	2.95	**-0.08**	0.84	9.30

sMAPE difference with baseline for multiple prediction lengths, L, for n=1000 for air dataset (1/3)

Confidence and Robustness

	Scale	Random	Shift	Offset	Min-Max
HTW	6.93	2.41	6.93	6.93	-0.58
EXP	0	2.84	0	0	3.38
KGW	-4	4.33	0.11	-4	6.17

z-score comparison for the baseline (no attack) and five self defined attack for multiple runs with the z-scores averaged

Conclusion

Research question answer:
Original Heads Tails Watermark algorithm serves as a robust and high quality watermarking method for time series foundation models

Limitations

- Time series foundation models are novel so quality retention and confidence performance may change for future models
- The watermarks have only been evaluated for a selection of attacks and could be vulnerable to other high-level attacks such as a Discrete Wavelet Transformation

Responsible Professor: Dr. Lydia Y. Chen (y.chen-10@tudelft.nl) , Supervisors: Jeroen M. Galjaard (J.M.Galjaard@tudelft.nl), Chaoyi Zhu (C.Zhu-2@tudelft.nl)