Multivariate Time Series Anomaly Detection based on Pre-trained Models with Dual-Attention Mechanism

Yongqian Sun[1], Yang Guo[1], Minghan Liang[1], Xidao Wen[3],
Junhua Kuang[1], Shenglin Zhang[1]*, Hongbo Li[2], Kaixu Xia[2] and Dan Pei[4]

Background

Design

Evaluation

Conclusion

Unavoidable Failures

- On April 8, 2024[1], Tencent Cloud experienced a major outage, rendering the console completely unavailable, causing a substantial negative impact on the reputation of Tencent Cloud services.

- On March 5, 2024[2], Meta services, including Facebook, Instagram, and others, experienced a disruption that prevented users from accessing those apps.

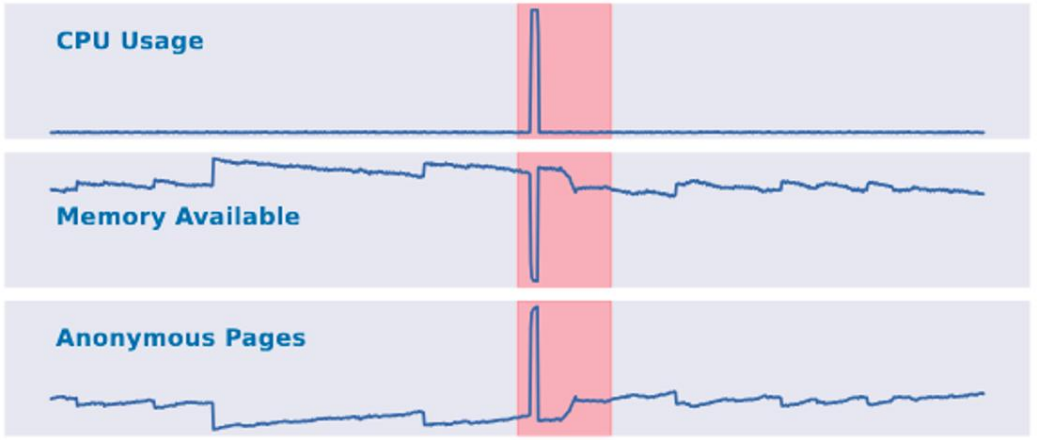
- OpenAI indicated ChatGPT experienced a major outage on June 4, 2024[3]. It recovered for a few hours, then went down again for about three hours.

Multivariate Time Series (MTS) Anomaly Detection

Part of the collected metrics.

Cache Penetration.

Part of the collected metrics.

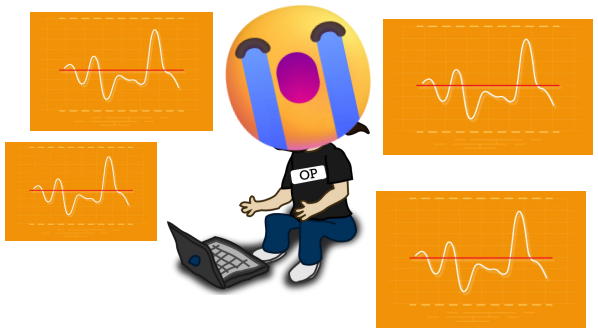


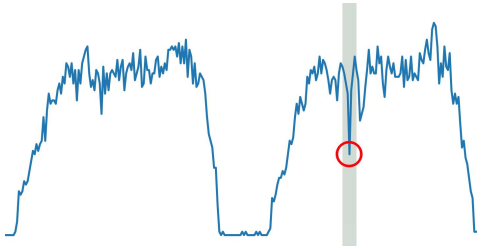
Request surge.

Therefore, anomaly detection based on monitoring MTS is vital important.

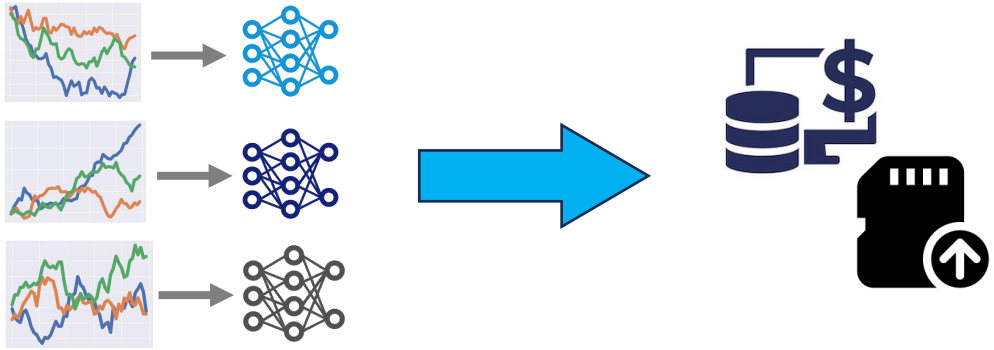
Previous Work

- **Traditional methods:** static rules
  - Requiring significant expertise and effort
  - Can not handle contextual anomalies



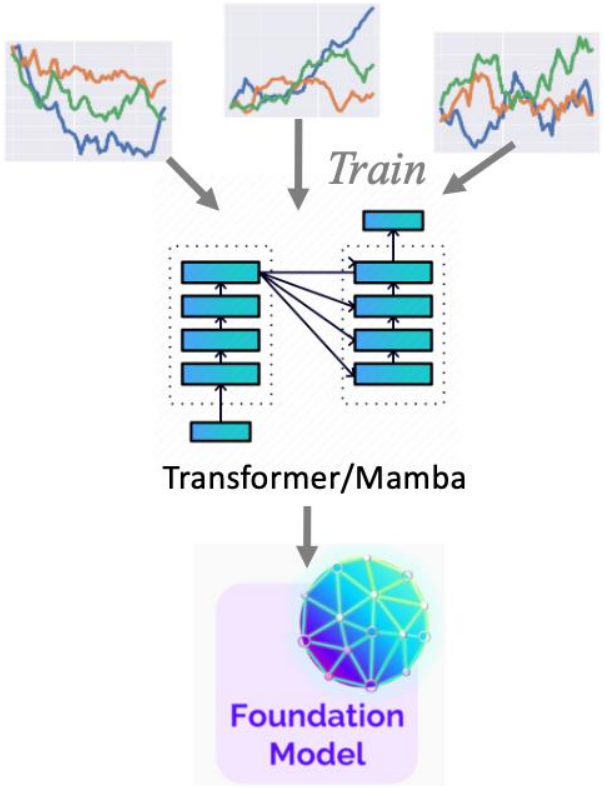


- **Popular methods:** use structures like VAE[4], RNN[5]
  - One model for each MTS or cluster
    Increased storage, memory costs

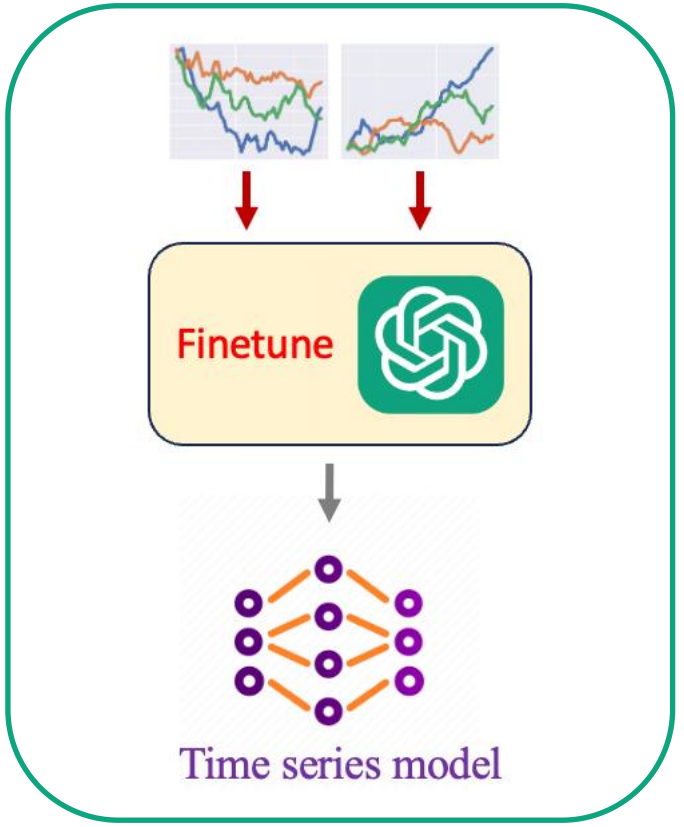


We need a method with strong generalization capability that can perform well in specific domain data.

Previous Work: Pre-trained models

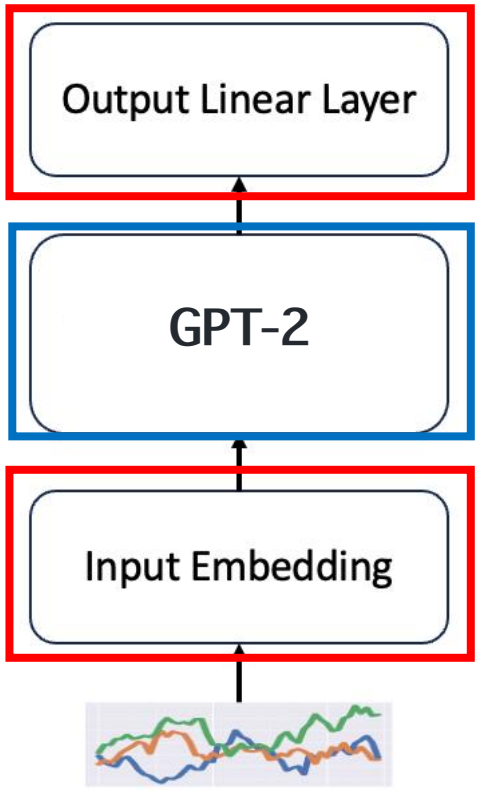


Solution 1:
Directly train a TS foundation model.



Solution 2:
Finetune a language model.
(e.g. *One-Fits-All* [6])

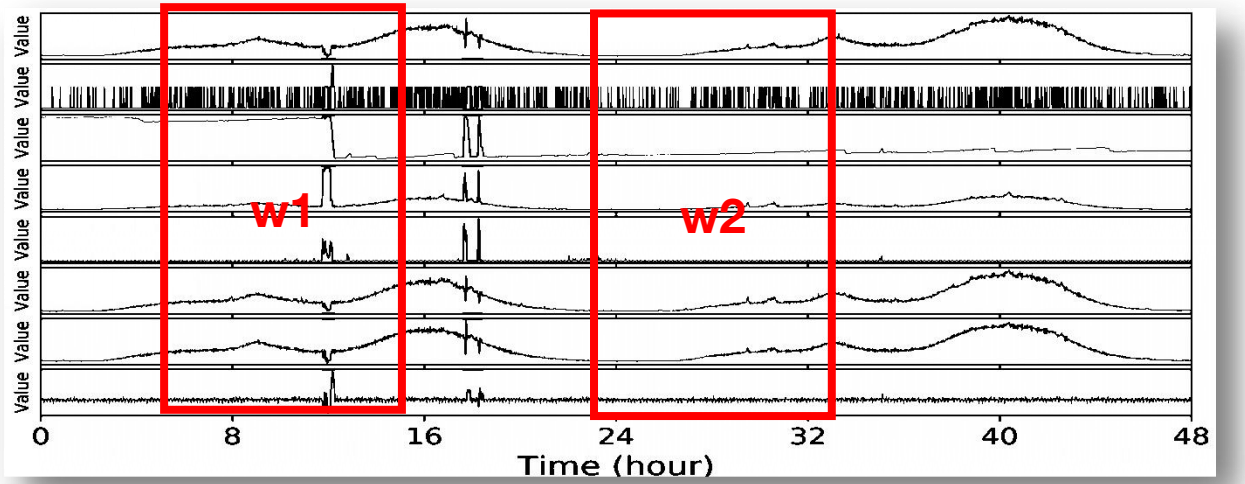
Previous Work: *One-Fits-All* [6]



Overview of *One-Fits-All* [6]

Shortcomings:

- **Challenge 1:** It can capture the temporal characters, but can not catch the inter-metric relationships very well for MTS.
- **Challenge 2:** Fine-tuning the GPT-2 at the same time with the input and output layer will cause degradation of GPT-2.



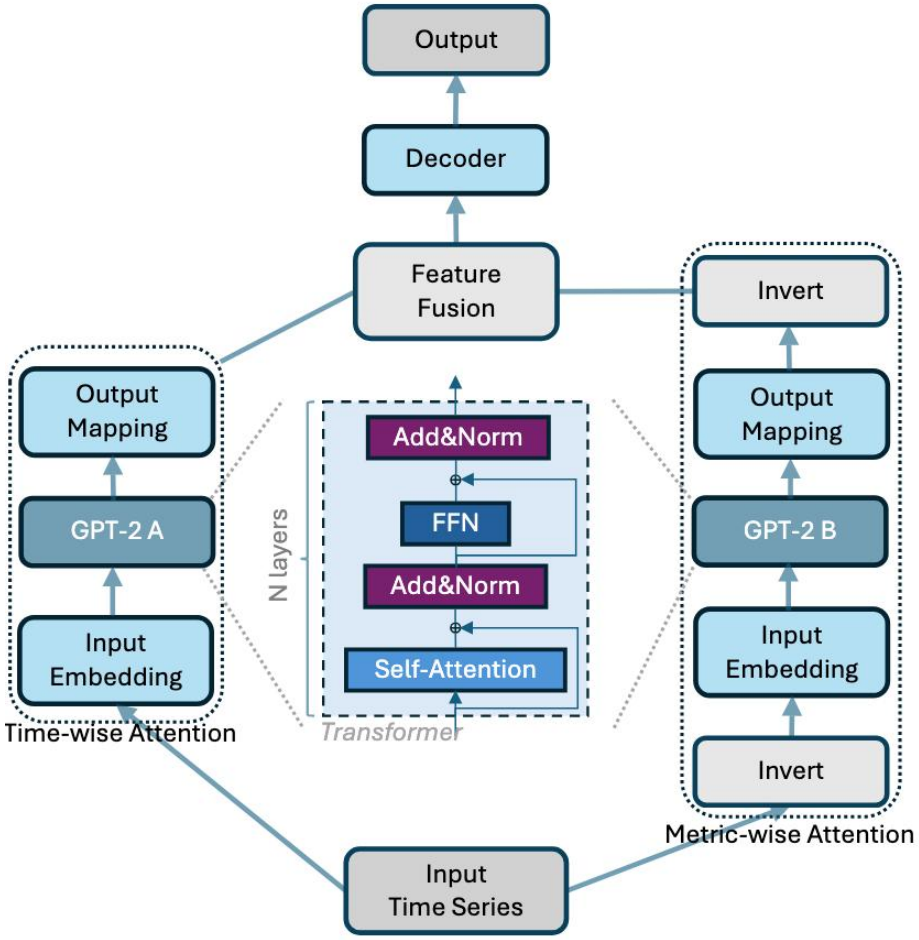
Background

Design

Evaluation

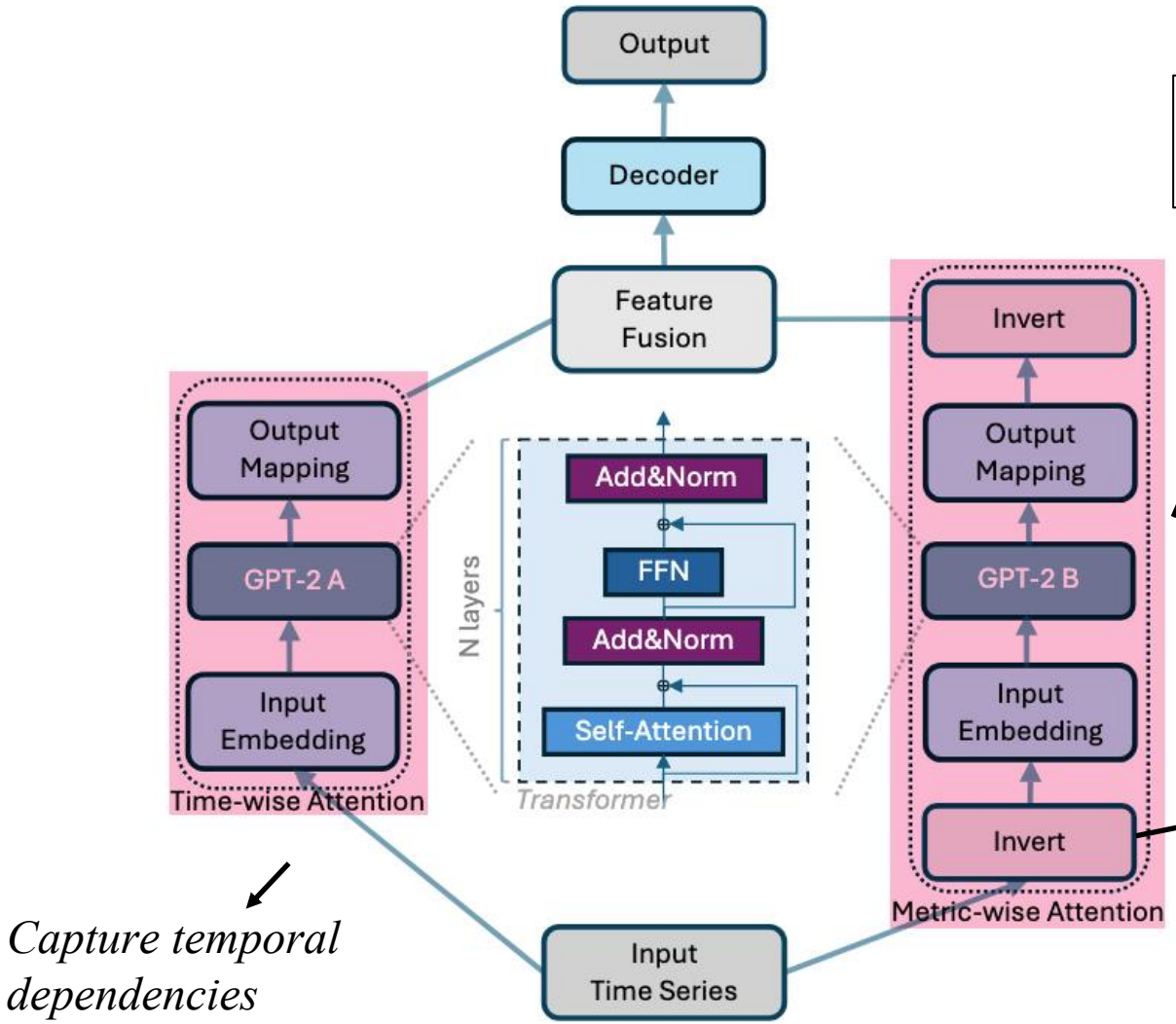
Conclusion

Overview



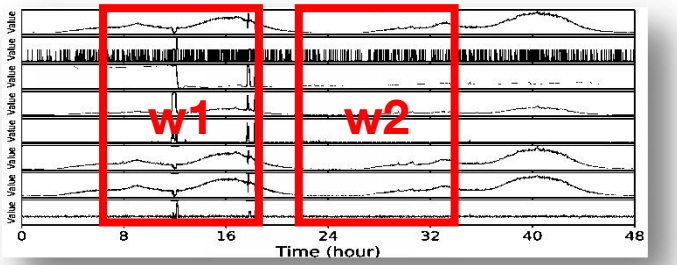
Overview of *DualLMAD*.

Dual Attention Mechanism (For Challenge 1)

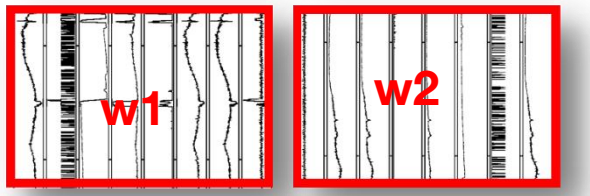


Challenge 1: **One-fits-all can not catch the inter-metric relationships well**

Capture inter-metric dependencies.



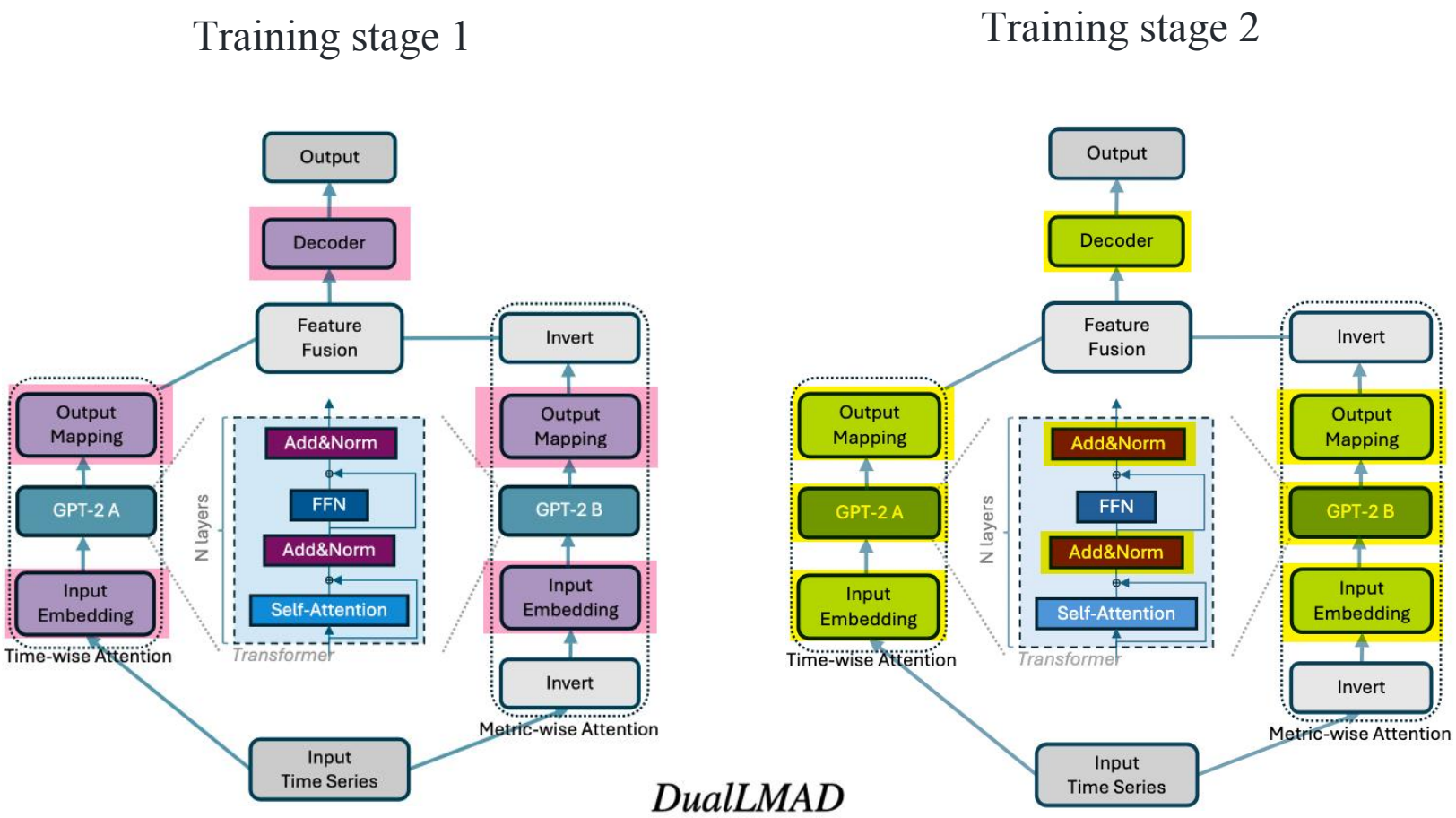
Invert



Capture temporal dependencies

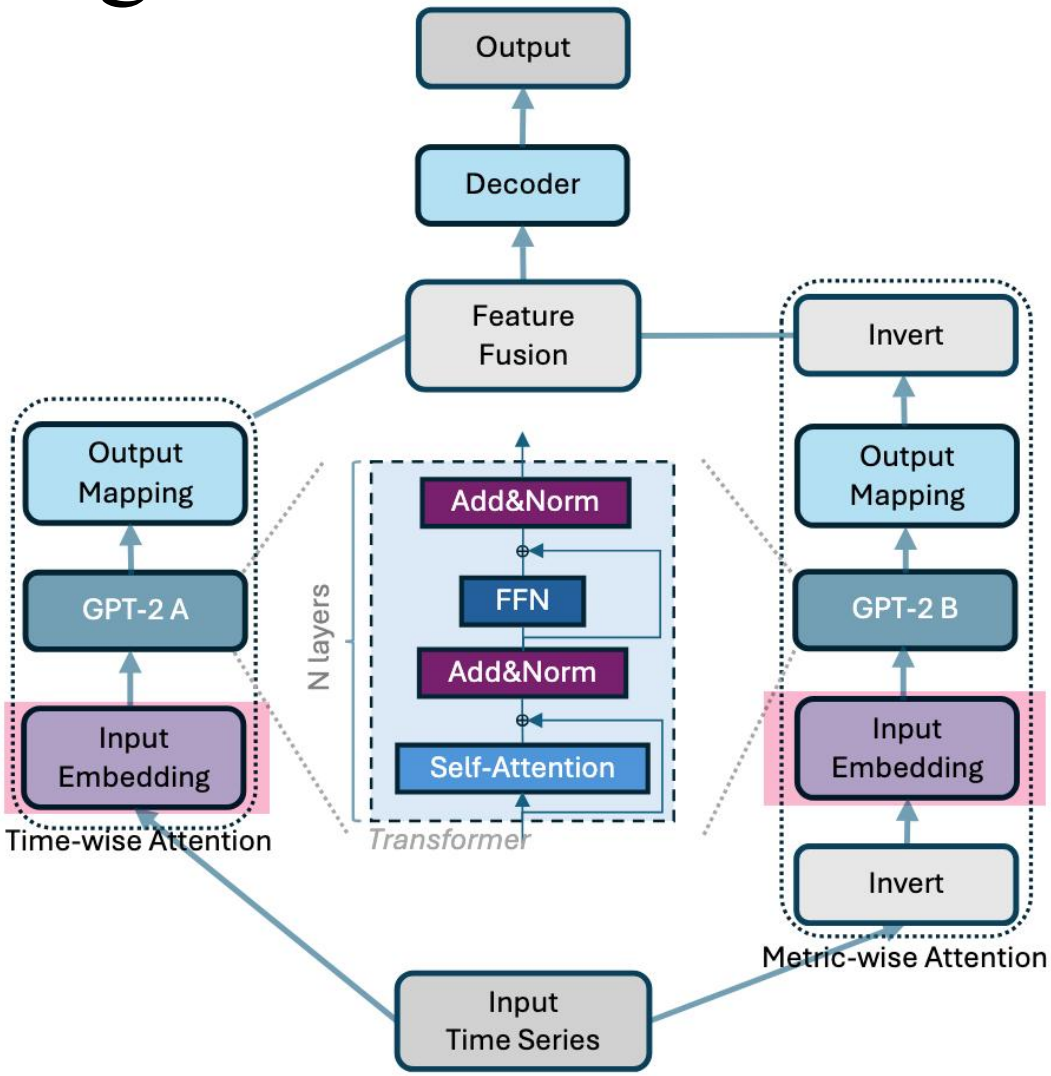
DualLMAD

Staged training strategy (For Challenge 2)



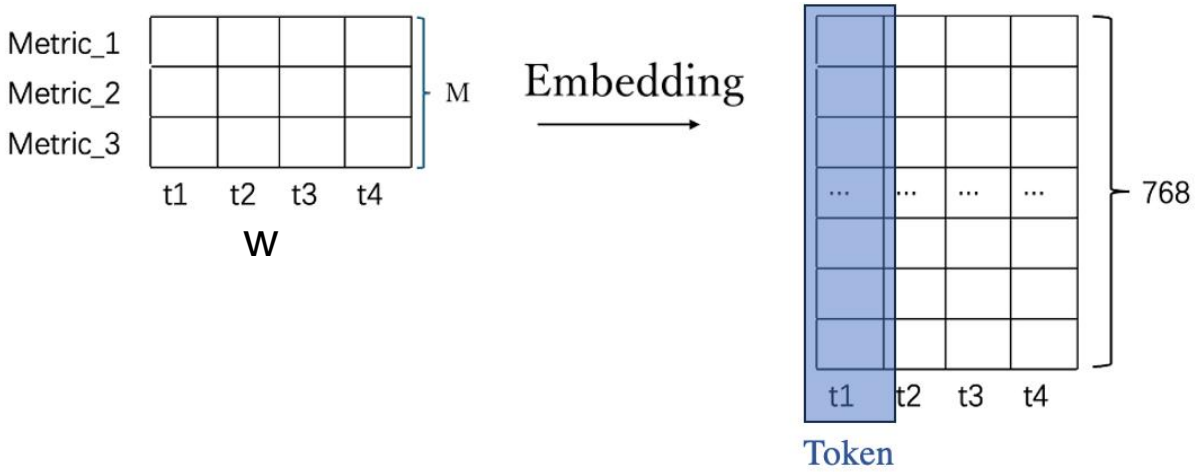
Challenge 2: Fine-tuning the GPT-2 at the same time with the input and output layer will cause degradation of GPT-2.

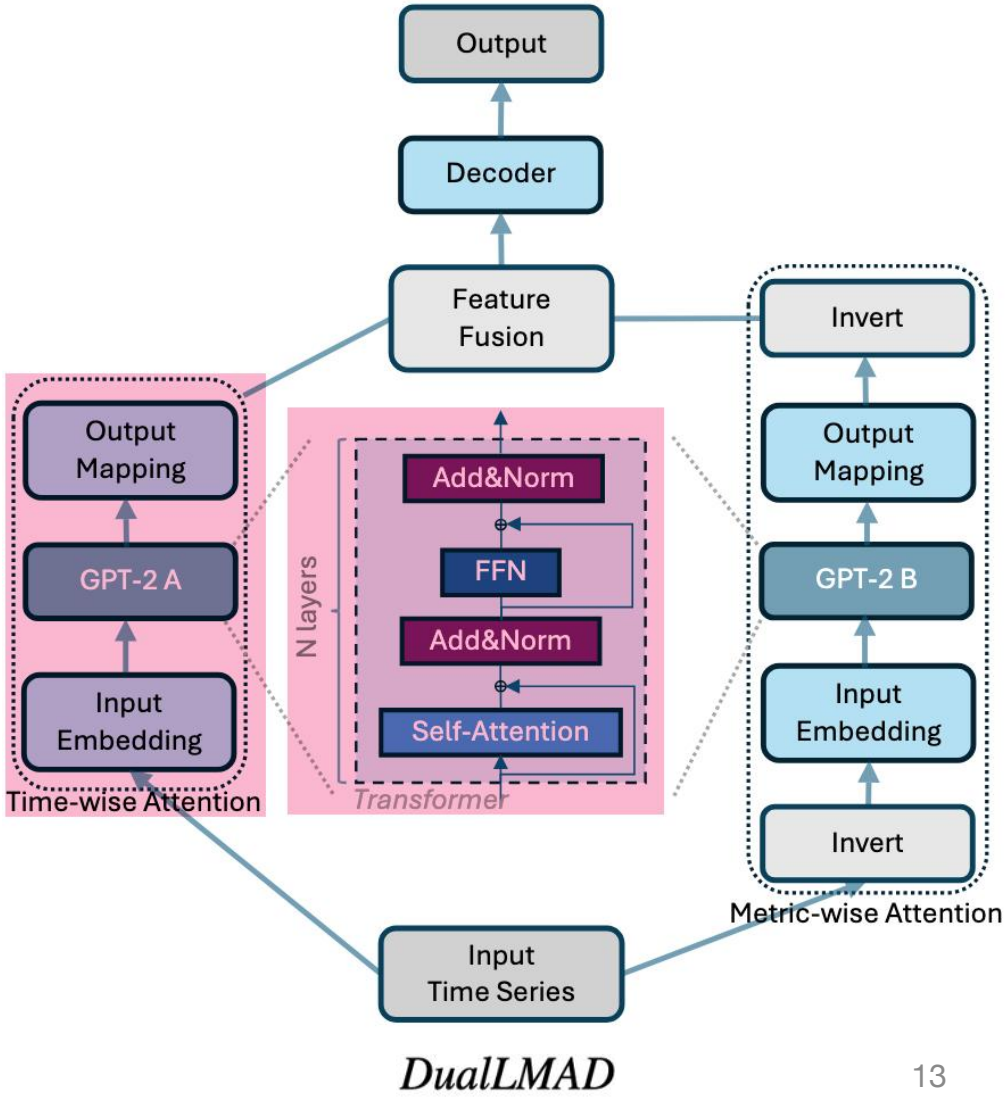
Input Embedding



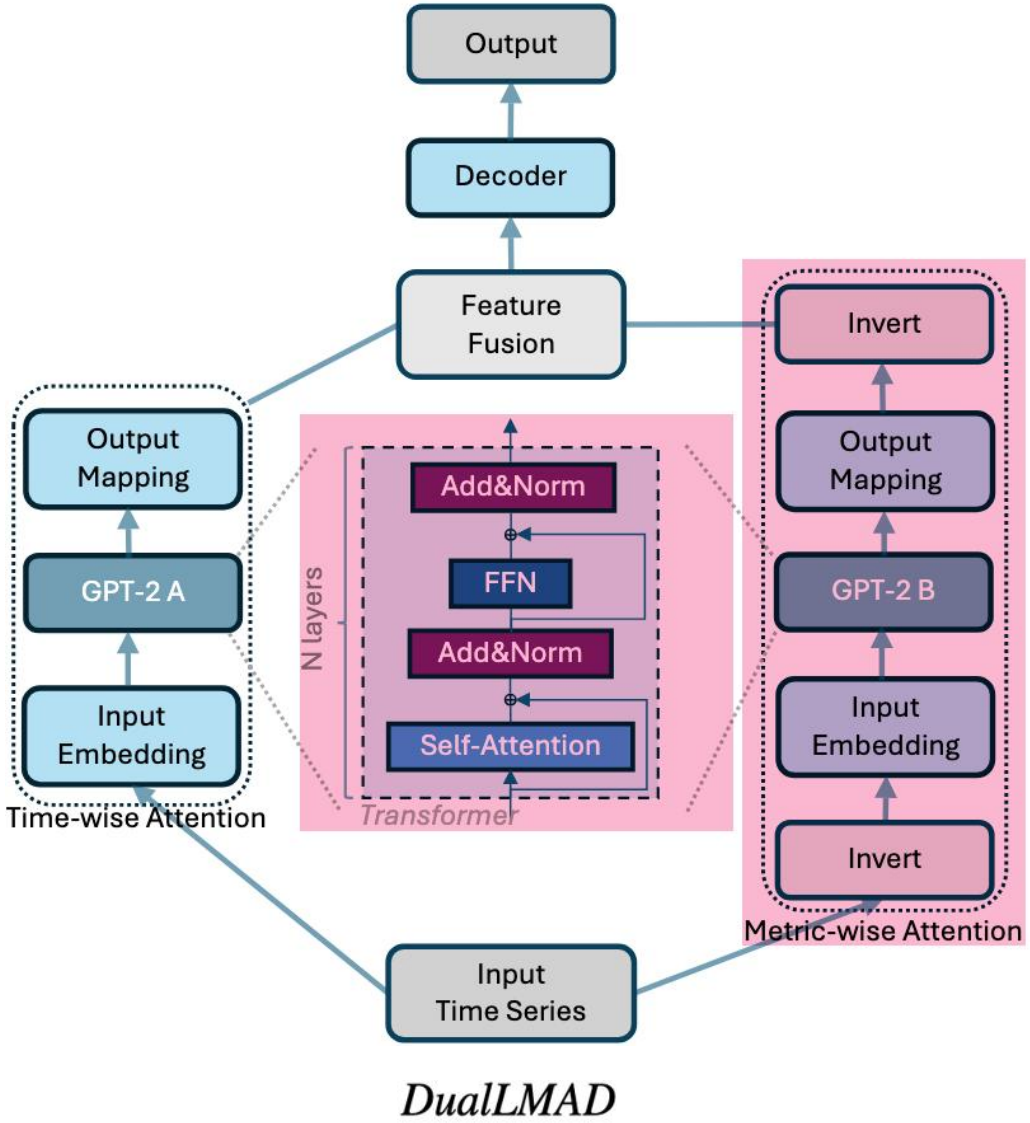
DualLMAD

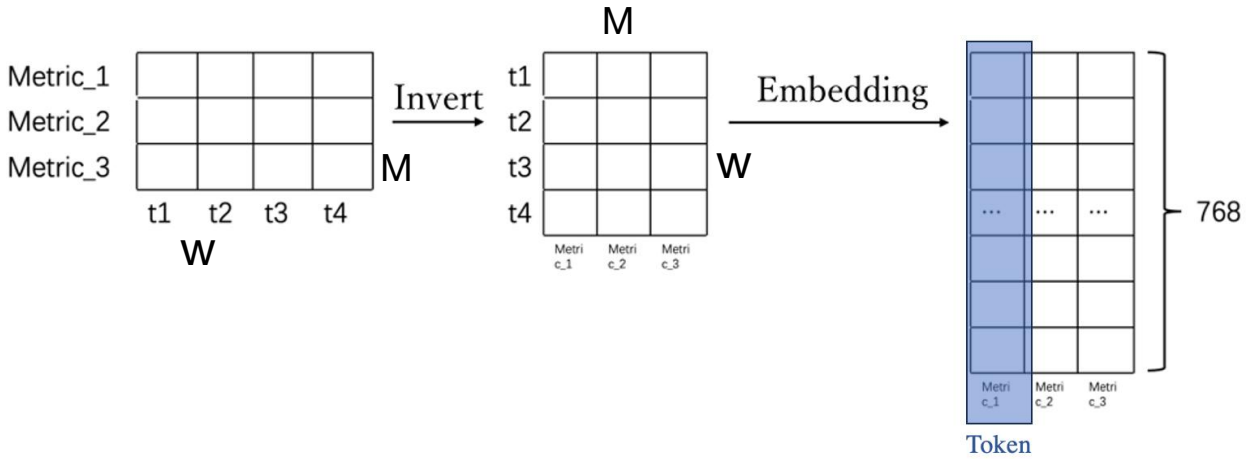
Time-wise self attention



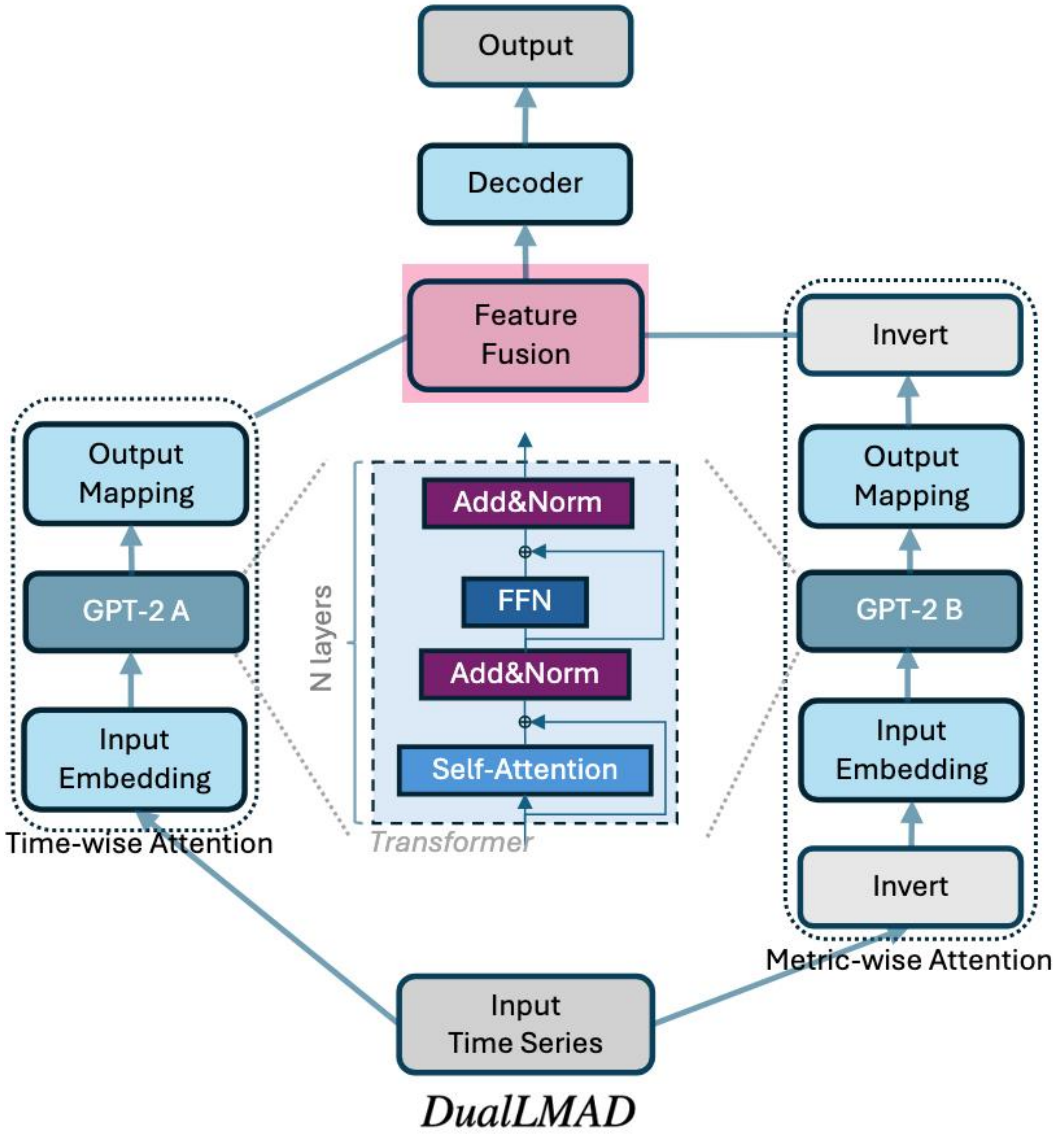


Metric-wise self attention



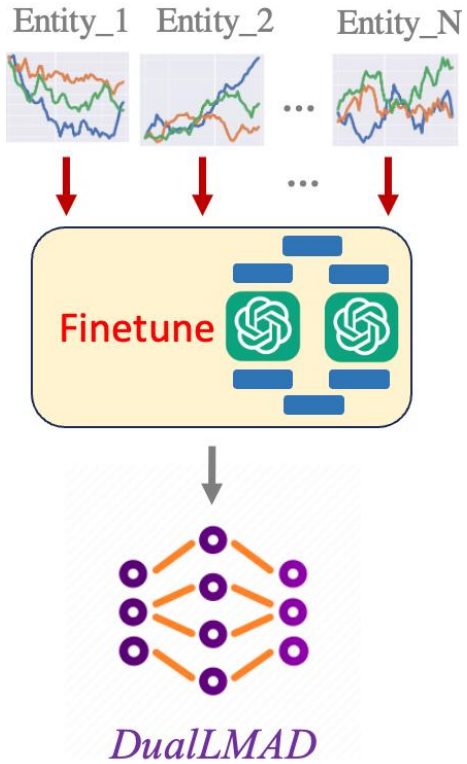


Feature fusion

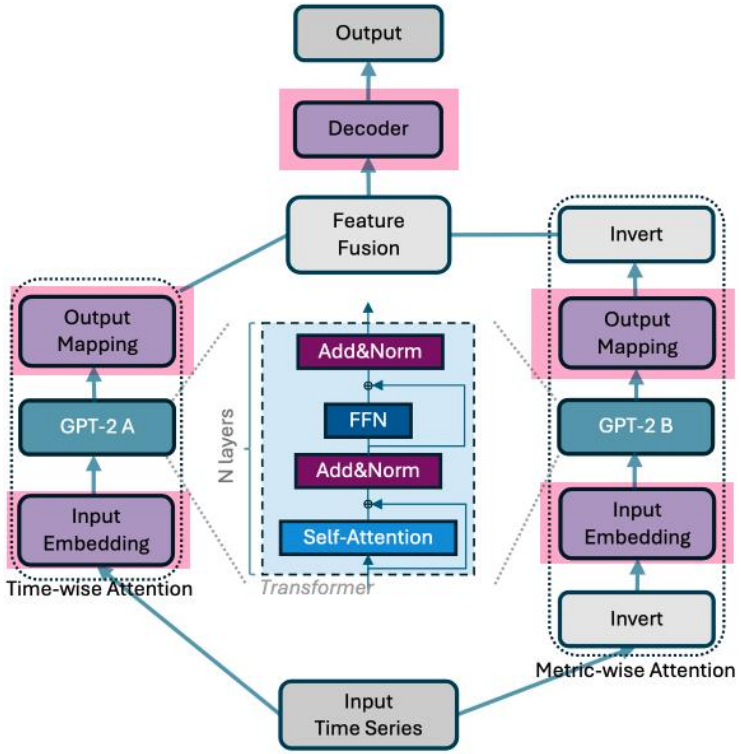


Model Training

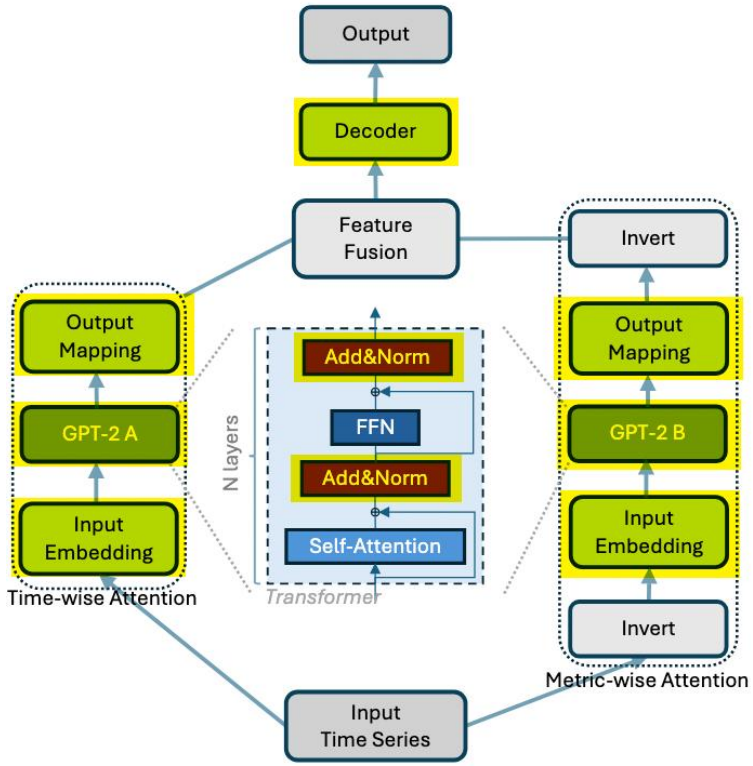
Train one model using data from multiple entities.



Training stage 1



Training stage 2



Background

Design

Evaluation

Conclusion

Dataset

Dataset Information.

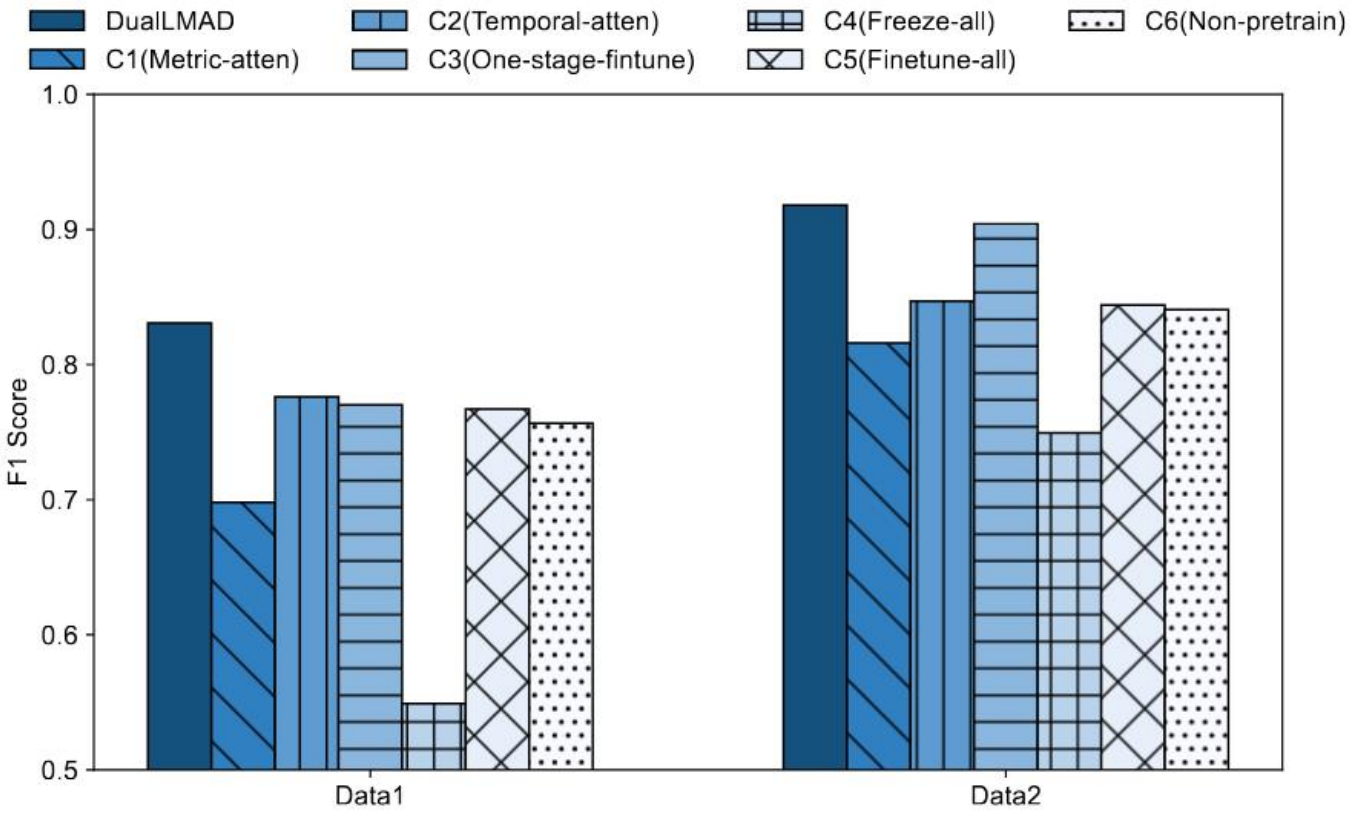
Dataset	Source	#entities	#metrics	Time points Train	Time points Test
SMD[5]	Internet Company	28	38	708405	708420
SMAP[7]	NASA	55	25	135183	427617
MSL[7]	NASA	27	55	58317	737729
Data1[7]	Global content service provider	200	19	134400	672*200
Data2[7]	Network supplier	200	25	28800	576*200

Overall Performance (share-model scenarios)

Share-model scenarios: Train only one model using all entity data from the training set to perform anomaly detection on the test set. For example, we only train one model on SMD dataset.

Method	SMD	SMAP	MSL	Data1	Data2
One-Fits-All[6]	0.8481	0.6887	0.8415	0.648	0.7511
TimesNet[8]	0.8457	0.6972	0.8184	0.7004	0.8146
iTransformer[9]	0.7119	0.6935	0.7254	0.6081	0.7793
USAD[4]	0.7892	0.6994	**0.8849**	0.2611	0.3653
OmniAnomaly[5]	0.6233	0.7036	0.8257	0.1767	0.3768
DualLMAD	**0.8661**	**0.7246**	0.8739	**0.8308**	**0.9180**

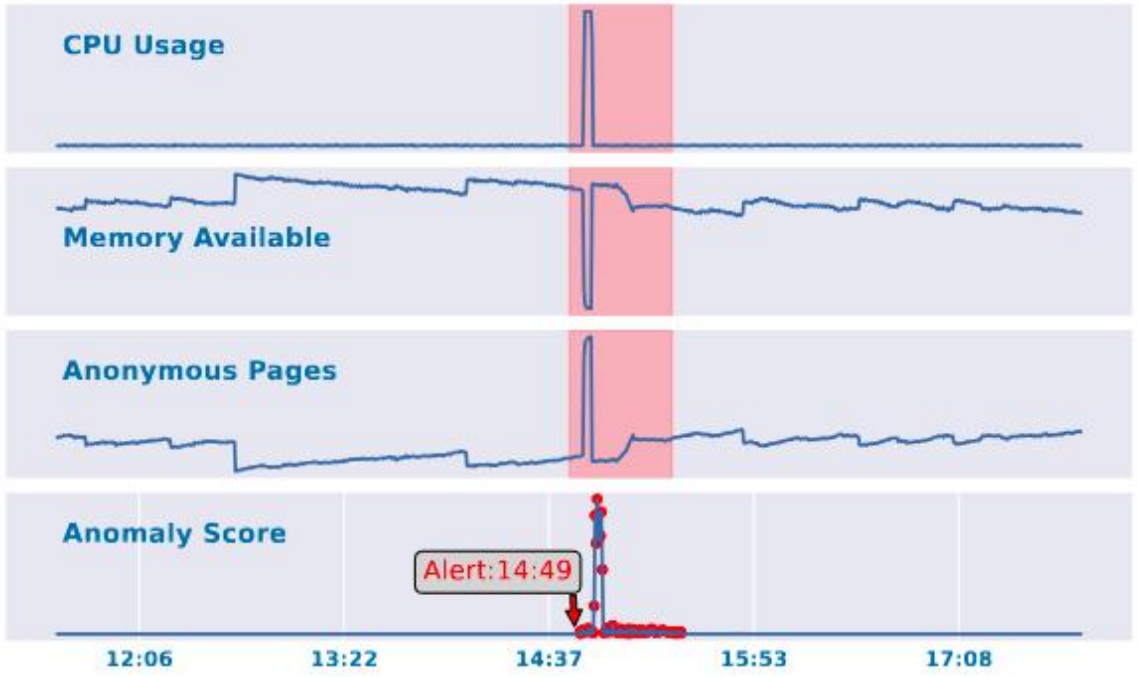
Ablation Study



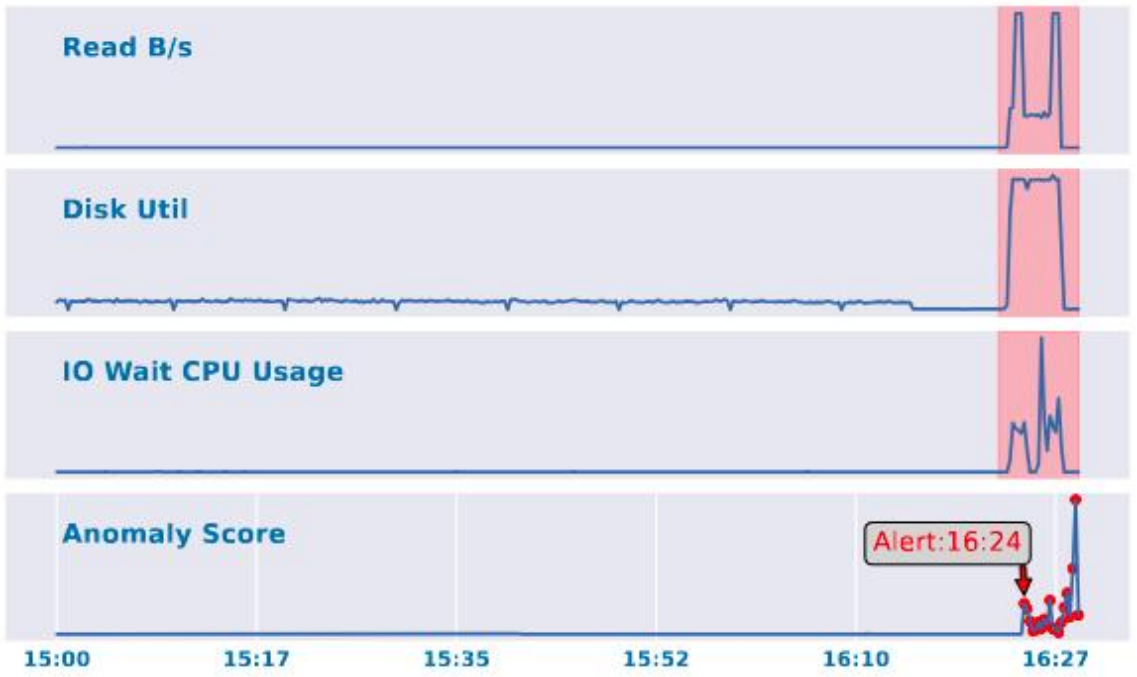
Performance of *DualLMAD* and Its Variants.

- C1 uses a single GPT-2 model to capture inter-metric relationships.
- C2 uses a single GPT-2 model to capture temporal relationships.
- C3 freezes the self-attention and FFN layers, performing a single-stage fine-tuning.
- C4 freezes all parameters in the pre-trained models.
- C5 fine-tunes all parameters in the pre-trained models during the second training stage.
- C6 randomly initializes the parameters in the pre-trained models.

Case Study



NoSQL Storage Business Cyber Exercise. *DualLMAD* detected the anomaly by identifying significant deviations in memory and CPU metrics.



Cache Penetration. After improper memory recycling strategies led to cache penetration risks, *DualLMAD* enabled operations personnel to quickly identify and address the issue.

Background

Design

Evaluation

Conclusion

Conclusion

- **We introduce a novel approach, *DualLMAD***, which leverages pre-trained language models for anomaly detection in time series data. By addressing the distinct characteristics of MTS, such as temporal and inter-metric relation-ships, through a dual attention mechanism, *DualLMAD* significantly improves generalization and accuracy.
- **We propose a staged training strategy.** In the first stage, only the input and output layers are updated. In the second stage, we fine-tune the add & norm layers, enhancing the model's ability to handle diverse time series datasets.
- **Extensive experiments** comparing with popular anomaly detection and time series algorithms across multiple datasets show that *DualLMAD* achieves superior performance. Additionally, we developed and validated an internal anomaly detection framework at a globally renowned internet company, demonstrating that *DualLMAD* meets the company's requirements for MTS anomaly detection.

References

[1] TencentCloud, “腾讯云4.8号重大故障复盘,” April 17, 2024.[Online]. Available: https://cloud.tencent.com/developer/article/2409985

[2] thousandeyes, “Meta outage analysis: March 5, 2024,” 2024.[Online]. Available: https://www.thousandeyes.com/blog/meta-outage-analysis-march-5-2024

[3] OpenAI, “06-04-2024-chatgpt-widespread-outage,” 2024. [Online]. Available: https://community.openai.com/t/06-04-2024-chatgpt-widespread-outage/801592r

[4] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A.Zuluaga, “USAD: unsupervised anomaly detection on multivariate time series,” in KDD ’20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, R. Gupta, Y. Liu, J. Tang, and B. A. Prakash, Eds. ACM, 2020, pp. 3395–3404. [Online]. Available: https://doi.org/10.1145/3394486.3403392

[5] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, “Robust anomaly detection for multivariate time series through stochastic recurrent neural network,” in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019, A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis, Eds. ACM, 2019, pp. 2828–2837. [Online]. Available: https://doi.org/10.1145/3292500.3330672

[6] Zhou T, Niu P, Sun L, et al. One fits all: Power general time series analysis by pretrained lm[J]. Advances in neural information processing systems, 2023, 36: 43322-43355.

[7] Li D, Zhang S, Sun Y, et al. An Empirical Analysis of Anomaly Detection Methods for Multivariate Time Series[C]//2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE). IEEE, 2023: 57-68.

[8] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, “Timesnet: Temporal 2d-variation modeling for general time series analysis,” in The eleventh international conference on learning representations, 2022.

[9] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, “itransformer: Inverted transformers are effective for time series forecasting,” arXiv preprint arXiv:2310.06625, 2023.

Thanks

sunyongqian@nankai.edu.cn

Q&A

Q1:Is it possible to replace GPT-2 with other pre-trained language models?
A1:We tried replacing GPT-2 with Llama-2-7B in the model, it still work.

Q2: The scale of deployment of this algorithm in partner companies is what, and what businesses is it used for?
A2:Due to the project deadline of the collaboration, approximately 400 machines was deployed for detection by the project's end. Specific business details cannot be disclosed due to confidentiality reasons.

Q3:What is the detection efficiency of this model?
A3:Due to the constraints on the length of the industry track, we did not include efficiency experiments in the paper. The efficiency of this model is slightly slower than that of GPT4TS and TimesNet, but it is still on the same order of magnitude.