

## Review 2

We would like to thank the reviewer for their efforts in improving our manuscript. We have edited the text in response to the reviewer's comments as follows.

The manuscript is primarily pedagogical, aiming at interpretability and transparency. While this is valuable, it does not propose fundamentally new architectures or theoretical insights.

You got correctly the primary objective of the paper, which is about reproducibility, transparency and interpretability. Our contribution does not represent another improvement to the basic framework, but rather a better understanding of the core mechanisms of fundamental processes within the core structure. To the best of our knowledge, transformers have not previously been described in complete pseudocode, nor have their memory requirements been detailed for each individual parameter to be learnt. Furthermore, even the simplest vanilla time series forecasting architectures available for download include some optimized details, whereas our code is simply the most basic implementation of the architecture proposed in Vaswani et al. (arXiv, 2017). We believe that our contribution is valuable to the research community because it provides an unquestionable baseline against which to validate the results of more elaborate additions. We have tried to make these points clearer in the edited manuscript, where all edited sections are in blue.

Finally, we note that the contributions are in line with the journal's aims and scope, stating: 'The aim of Algorithms is to encourage scientists to publish their experimental and theoretical results'. Our contribution is primarily experimental, with theoretical value as indicated above.

Compared with recent transformer adaptations (Informer, Autoformer, FEDformer, PatchTST), the novelty is limited.

In addition to the points outlined above, we now also provide a computational comparison against the results obtained using two advanced time series forecasting architectures. This allows us to evaluate the difference in quality between the results obtained using our minimalist code and those at the forefront of the field.

Mathematics in the article is largely a restatement of known transformer operations.

It is our intention to use exactly the mathematics proposed in Vaswani et al. (Arxiv, 2017). We now mention this more explicitly in the text.

The “minimalist” approach does not introduce new theoretical insights.

See the answers above. We hope that the current version of the text makes our position clearer.

1. The title is too generic; the authors must have to modify the title to be specifically linked with the proposed contribution.

We updated the title.

2. The abstract repeatedly uses the pronoun “we” (i.e., “we describe,” “we implement,” “we validate”). I recommend rephrasing these sentences in the passive voice or by referring directly to the work. The same goes for the conclusion, too.

We updated both abstract and conclusions as requested.

3. Clearly articulate the study's contributions and novelty, supported by key findings/quantitative results for abstract enhancement.

We included more explicit sentences stating the contributions, which are aligned with the above answers. We also significantly expanded the quantitative results that support our claims.

4. Evaluation is limited to univariate series; multivariate testing would provide stronger validation. Could the minimalist approach be extended to multivariate data, which is more realistic in forecasting applications?

We mentioned this possibility in the conclusions.

5. A block diagram of the overall architecture should be included. Adding a visual representation of the transformer-based model will improve clarity and help readers better understand the workflow and component interactions.

We added the diagram as a further figure.

6. Hyperparameter tuning procedures are not sufficiently detailed. Results may depend strongly on initialization, learning rate, and sequence length choices. How robust is the model to missing data, noise, or seasonality beyond simple normalization?

We included a description of how we optimized the hyperparameters. This was achieved using the Optuna package for the transformer and the enhanced grid search built into the autoregressive models. The resilience against unwieldy real-world time series was – and is – exactly one of the objectives of Makridakis contests, among which we chose to use the M3 data. This guarantees to be able to cope with seasonal and noisy data as are commonly found in business forecasting. We added a comment to this effect in the text.

7. Statistical validation (Mann-Whitney U-test) is appropriate, but the depth of interpretation is minimal.

We added a further comment besides simply commenting on the significance or not of the tests.

8. Performance differences across domains are reported but not deeply analyzed (e.g., why finance performed better than demographics).

We included a tentative interpretation of the reasons of the observed results.

9. Scalability test is not included for the proposed scheme (e.g., long-horizon forecasts on very large datasets).

We added results obtained on longer, real-world series generated by IoT sensors. We also included a comment stating that in the context of univariate time series forecasting, extending the historical window too far back risks incorporating data generated under structural regimes that are no longer in effect. Conversely, extending the forecast horizon too far ahead is of limited utility as predictive accuracy inevitably deteriorates with temporal distance. Therefore, although longer series exist, they are of limited interest to our discussion. This is evidenced by more recent forecasting contests that use series of a similar size to those proposed when testing effectiveness on long time series.

10. Discuss the trade-offs between interpretability and accuracy more explicitly.

We have included a new subsection, 4.4, which is dedicated to the discussion of the interpretability of attention data. Accuracy, as related to model dimension, is discussed in the new subsection 4.5.

11. Include an ablation study showing how much accuracy is lost at each simplification step compared to a standard transformer.

We have included a dedicated subsection: 4.3.

12. It is recommended to enhance Figures 2, 3, 4 & 5 and their font sizes, as these are not properly readable.

We have increased the font size of all the figures, including the ones added for the revision.

13. Computational efficiency (training time, memory footprint, scalability) is not analyzed. Since the model is advertised as “minimalist,” these aspects should be emphasized.

We now report training times in the tables, while memory footprints are reported by detailing the dimensions of all data structures constructed during processing. This applies to both the tiny restaurant example and the larger airline passengers and M3 examples.

14. Strengthen the discussion on computational efficiency (training time, parameter count vs. accuracy).

We have included a dedicated subsection: 4.5.

15. Benchmarking is limited to Random Forest, which is a relatively weak baseline. Comparisons with ARIMA/ETS and recent transformer variants would provide a more meaningful context.

We have expanded the comparison subsection considerably by including comparative results against ARIMA, Holt and Winters (as representative of ETS) and two state-of-the-art transformer architectures specialized for time series forecasting. These results are reported in Table 4.

16. A few areas for improvement are needed for English. Some sentences are overly long and could be made more concise (e.g., in the Introduction and Methodology). Some redundancy in stating that the architecture is “minimalist” and “transparent” multiple times.

We checked the text, aiming to reduce each sentence to a maximum of two lines and ensuring that no sentence exceeded three lines.