

Peer Review

Review of: "LLM-ABBA: Understanding Time Series via Symbolic Approximation"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The authors take ABBA, an existing symbolic approximation that compresses a time series into a polygonal chain and then clusters the (length, increment) pieces into a small alphabet, and feed the resulting strings to language models fine-tuned with QLoRA. The symbols are drawn from the model's own vocabulary, so no new embedding layer is trained. Three backbones are used (RoBERTa-large, Llama-2-7B, Mistral-7B) on three task families: classification on 21 UCR datasets (Table V) and three medical sets, EEG eye state, PTB-DB, and MIT-BIH (Table VI); extrinsic regression on the 19 TSER datasets (Table VII); and forecasting on the four ETT subsets (Table VIII). On the method side, the paper adds a fixed-point variant of the compression step (FAPCA, Section III.C.1), meant to stop one wrongly predicted symbol from shifting the whole reconstructed tail, plus a joint symbolization scheme for multiple series and for out-of-sample data (Sections III.C.2 and III.C.3). Section III.B gives error bounds for the reconstruction (Theorems III.1 to III.5).

The headline claims are a new state of the art on TSER, classification that "compares favorably" to SOTA on UCR and the medical sets, and forecasting that is "competitive" with recent methods. There is real value here. The idea of reusing pretrained tokens as an ABBA alphabet is simple and clever, the regression results for RoBERTa-large are strong, the FAPCA fix addresses a genuine weakness of symbolic forecasting, and the code is public. My concerns are mostly about the distance between what the tables show and how the abstract and discussion describe them, and about several numbers in Tables III, VII, and VIII that cannot be right as printed.

2. Whether the conclusion holds on the data

I start with what holds. The regression result is the most convincing part of the paper. Comparing the best RoBERTa-large-ABBA value across the three ranks with the better of DrCIF and FreshPRINCE in Table VII, I count RoBERTa ahead on 16 rows, behind on 2 (HouseholdPowerConsumption1: 377.02 against 146.96; BenzeneConcentration: 4.00 against 3.57), with AustraliaRainfall having no baseline values. Interestingly, that is better than the text itself says: Section IV.E reports "12 out of 19 use-cases". I would reconcile the two, and if the 12 refers to a comparison against the 20 regressors from [46] in Figure 7, say so. The average rank of 2.72 with a critical difference of 7.356 (Section IV.E) supports "best average rank", but with a CD that wide, RoBERTa-large-ABBA is not significantly separated from any regressor ranked within about 7.4 of it; I would state that alongside the ranking.

Classification is where the abstract overreaches. In Table V, V2S (column SOTA) is ahead of every LLM-ABBA variant on 20 of the 21 datasets; the exception is Earthquakes, where Mistral-7B-ABBA with k-means (J1) reaches 79.1 against 78.4. Elsewhere, the gaps are often large: FordA, best LLM-ABBA 68.9 (RoBERTa, J1 and J2) against 100; TwoLeadECG, 69.1 (RoBERTa, J2) against 97.8; ItalyPowerDemand, 73.4 (Llama-2, J2) against 97.1. Apart from Earthquakes, only Herring, Wafer, and Coffee come within 4 points. Section IV.D says as much ("LLM-ABBA cannot outperform the SOTA"), yet the same paragraph ends with "achieves a noticeable improvement", and a later paragraph of Section IV.D speaks of "only a small performance gap" and results "within a narrow margin of the current best". Those sentences and the abstract's "compares favorably" do not match Table V. The medical results in Table VI are fairer to the claim: on ptb-db, Llama-2-7B-ABBA reaches 99.0 (r=16) against CNN 99.4, and on MIT-BIH, the best value, 89.7 (Mistral, r=64), is above LSTM 88.1 but below CNN 93.4 and BiRNN 96.5. That is honestly described in the text.

For forecasting, Section IV.F says LLM-ABBA "compares favorably to the Informer architecture." I checked every cell of Table VIII where Informer is reported (six rows, horizons 24 and 168). In all six, Informer has the lower MSE, and in five of them also the lower MAE; the one exception is ETTh1 168/24, where Llama-2-7B-ABBA (r=64) reaches MAE 0.364 against 0.369. Two examples: ETTh1, 168/24, best LLM-ABBA MSE 0.622 (Mistral, r=64) against Informer 0.577; ETTh1, 168/168, best MSE 0.966 (Mistral, r=64) against Informer 0.678. Against Time-LLM and TimeMixer at horizon 96, the difference is roughly a factor of two (ETTh1 168/96: LLM-ABBA MSE between 0.754 and 0.792, Time-LLM 0.362, TimeMixer 0.375). So the forecasting sentence in Section IV.F, the phrase "comparable time series forecasting results to the SOTAs" in Section IV.G, and "competitive forecasting capability" in the abstract should all be softened.

Now the numbers that cannot be right as printed. For any set of errors, RMSE is at least MAE, so MAE can never exceed the square root of MSE; this also survives averaging over channels or windows. Table VIII breaks this repeatedly. ETTh2, 168/168, Llama-2-7B-ABBA $r=16$: MSE 4.103 gives $\sqrt{4.103} = 2.026$, yet MAE is 2.675. All six LLM-ABBA cells of that row violate it (for Mistral $r=64$: $\sqrt{3.898} = 1.974$ against MAE 2.134). ETTh1, 168/168, Mistral: $r=16$, $\sqrt{1.182} = 1.087$ against 1.217; $r=64$, $\sqrt{1.174} = 1.084$ against 1.968; $r=256$, $\sqrt{1.179} = 1.086$ against 1.992. The same issue appears in Table III, first row ($\text{tol} = \alpha = 10^{-2}$): MSE 2.5×10^{-7} gives an RMSE of 5.0×10^{-4} , well below the reported MAE of 1×10^{-2} . Either columns are swapped, the metrics are computed on different scales, or there is a computation error. Section IV.B also labels "mean-absolute-error (MSE), mean-square-error (MAE)" the wrong way round, which may be related.

Table VII has two suspected copy errors. The IEEEPPG baselines (DrCIF 34.92, FreshPRINCE 33.69) are identical to the LiveFuelMoistureContent baselines six rows above. And BIDMC32SpO2, NewsHeadlineSentiment, and NewsTitleSentiment all list 5537 symbols. I also wondered about the scale of some RMSEs: on FloodModeling1, RoBERTa gives 0.002, Llama-2 0.05, and Mistral 0.37, a spread of more than two orders of magnitude between backbones. It would help to confirm that all RMSEs, including those taken from [46], are computed on the original target scale.

Table IV (noise on MIT-BIH) reports 4,578 symbols at 0 dB, which is exactly the symbol count for clean MIT-BIH in Table VI, and 10,642 at 20 dB, i.e., more symbols with less noise. I would expect the opposite at fixed tol and α . Accuracy at 0 dB (89.2, $r=16$) is also higher than at 5 dB (87.2). Maybe there is a simple explanation. I could not find it in the text.

Finally, every result is a single run. No standard deviations are given, and the three LoRA ranks are the only source of variation. On UCR sets such as Coffee or Wine (test sets of a few dozen series), a one-run difference of a few points is within noise.

3. Reproducibility

The public repository (github.com/inEXASCALE/llm-abba) is a real strength, and Tables I and II give LoRA ranks, alpha, dropout, the inhibition threshold, learning rate, and batch size. The hardware (one A100 40GB) and 4-bit quantization are stated. As rendered on Qeios, the "Quant." and "Optims." columns named in the captions of Tables I and II are not legible.

What I could not find:

- Exact model checkpoints. "Llama-2-7B" and "Mistral-7B" exist in several releases (base and chat; Mistral v0.1, v0.2, v0.3). Hugging Face identifiers with revisions would settle it.
- Seeds, the exact number of epochs per dataset (Section IV says only "fewer than 10"), and the number of runs.
- How tol , α , and scl were chosen from the grids in Section IV.A.1, and on which split. UCR ships only train and test sets; if the grid was searched on test accuracy, the results are optimistic.
- Whether the ABBA symbol centres for forecasting and regression are fitted on training data only (Section III.C.3 describes the out-of-sample procedure, but Section IV does not say it was used).
- For forecasting: univariate or multivariate setting on ETT, the train/validation/test split, how a generated symbol string of variable total length is turned into exactly H points (step ⑦ of Figure 3), and whether Time-LLM and TimeMixer were rerun at input length 168 or copied from their papers (their papers mostly use longer inputs).
- The source of the EEG eye state baselines in Table VI. References [42], [43], and [44] are ECG papers; if the authors trained these models on EEG themselves, the training setup should be given.

4. What specifically to fix

1. Table VIII, rows ETTh2 168/168 (all six LLM-ABBA cells) and ETTh1 168/168 (the three Mistral cells): MAE exceeds $\sqrt{\text{MSE}}$, which is impossible. Please recheck the computation or column order, and do the same for Table III, first row. While there, fix the malformed cells in row ETTh1 168/96: "0. 60" (Llama-2, $r=256$, MAE) and "0.7811" (Mistral, $r=256$, MSE).
2. Abstract and Section IV.D: replace "compares favorably to the recent state-of-the-art (SOTA) in UCR" with a sentence that matches Table V, where V2S leads on 20 of 21 datasets, often by more than 10 points. The medical claim can stay, phrased as in Section IV.D.
3. Section IV.F: the Informer comparison. Informer has the lower MSE in all six rows where it is reported.
4. Section IV.E: "12 out of 19" does not match my count of 16 of 18 from Table VII. Also, add the CD caveat to the ranking claim.
5. Table VII, row IEEEPPG: the baseline values duplicate LiveFuelMoistureContent. Row AustraliaRainfall has empty baselines; say why.
6. Table IV: explain why the symbol count rises from 4,578 (0 dB) to 10,642 (20 dB), and why the 0 dB count equals the clean count in Table VI.

7. Table VIII header: Informer is cited as [2], which is the TSER paper by Tan et al.; the Informer reference is [48].
8. Report mean and standard deviation over at least three seeds for the main tables, or state clearly that all numbers are single runs.
9. Section III.B, "Modeling assumption": the local deviations are assumed independent, but Theorem III.2 forces their sum to (0,0). I would restate the assumption as independence conditional on that constraint (which is what a Brownian bridge is) and check that the Hoeffding step in the proof of Theorem III.5 still applies in that form.
10. Section III.D (Zipf): Figure 5 is only visual. A fitted exponent per dataset with a goodness-of-fit measure would turn "coarsely meet" into a result.
11. Section IV.J and Figure 10: the LLMs run on a GPU, the other regressors on a single CPU thread. That comparison favors LLM-ABBA and should be flagged in the text.
12. Smaller items: Table IX caption says "on ETTh1 data" but reports all four ETT subsets; "Mistreal" (Tables V to IX), "ReBERTalarge" (Sections IV.E, IV.J), "depdnds" (Table III caption), "epoches" (Section IV); Table V caption says "Full comparison" while the text calls it a partial set. Please also state how the 21 UCR datasets were selected. In the Figure 3 caption, as rendered on Qeios, step numbers repeat ("① and ①"), so the mapping to the figure is unclear.

5. Recommendation

Minor revisions. The method is original, the code is open, and the regression results in Table VII stand on their own. What needs work is a recheck of the forecasting and reconstruction metrics in Tables III and VIII, two copy errors in Table VII, and a rewrite of the abstract and Sections IV.D, IV.F, and IV.G so that the claims about classification and forecasting match the tables. None of this requires new ideas. I expect it can be done with the existing runs plus a few seeds.

Declarations

Potential competing interests: No potential competing interests to declare.