

Dear Anonymous Reviewer,

We appreciate the time and effort you put into the detailed review. We considered all comments and changed the paper as follows:

Main Issues

1. We intentionally use vector representations for our concepts rather than scalars (as done by [1]), because we want the concepts to be defined across the temporal dimension of the time series and find scalar concepts too restrictive. Vector-represented concepts are indeed not necessarily human-understandable. Unlike [1], we therefore do not aim to interpret the concept scores themselves. Instead, we claim that we can localize these concepts into model components. By using the AR activations as representations of the AR surrogate model, we teach the model to represent its own activations similar to AR. This is a key design choice and we made it more explicit in the introduction.
2. We removed the phrase “leave the model’s architecture intact”, and added the necessity of the minor architectural change (page 2). This change (removing the residual connection) can be executed for any transformer.
3. We performed the requested ablation experiment. We remove the residual connection at the attention or feed-forward bottleneck location, but omit the CKA loss (see Table 3 on page 13).
4. Intervention is not the main goal, instead it is a method to prove concept *localization*. Learning the hour-of-day information is indeed trivial given the input, but identifying where this information is encoded within the model is not. In particular, with our method and intervention, we show that: (1) hour-of-day information is encoded in a specific model component, and (2) this component is the sole representation used downstream for this information, as intervening on it alone restores the original performance.
5. The AR model is indeed trained to predict one future value at a time. To obtain the same long term forecast as Autoformer, it needs to be applied recursively, generating intermediate values used for obtaining the final value. That’s why we use the stacked values of the rolling input window weighted by its coefficients as AR’s “activations”. These values do not contain the bias term, but do contain the other model coefficients. We specified this in the Method section on page 7.
6. We narrowed the claim of Fig. 5.

7. (1) The synthetic dataset is deliberately simplified. Its purpose is to provide a controlled setting in which the ground-truth concepts are known, allowing us to test whether our framework can localize pre-defined concepts. Its purpose is not to simulate the difficulty of real forecasting problems. (2) The timestamps are generated from a deterministic daily sequence beginning at a fixed calendar date (6 March 2000), and the model receives derived temporal features such as hour-of-day, day-of-week, day-of-month, and day-of-year, but not the absolute sample index or year. Although these features are informative, they do not provide a direct lookup from timestamp to target, and the target is not recoverable from time alone. The concepts are defined by their underlying functions and their composition over the preceding input window, rather than by a trivial read-off of the timestamp.
8. Figure 6c presents the forecast for one single test sample, whereas the results in Table 2 are averaged over the full test set (containing 20% of the full dataset). Therefore, the performance shown in this individual forecast is not necessarily representative of the mean test-set performance reported in the table.
9. We made the claim in Fig. 10 more precise and added a discussion on why the intervention over attention CBTL barely helps. We added an experiment where we repeated the analysis for the vanilla Transformer model, and we find that there is no difference in effectiveness between the Att and FF CBTL. Therefore, we suspect that the difference between the Att and FF CBTL is due to the architectural changes of the attention mechanism of the Autoformer (by containing auto-correlation). We suspect that the concept of time is represented in a more complex manner in this set-up, which results in more downstream effects.
10. We agree that the AR model achieves lower errors on four of the six benchmarks and is more inherently interpretable. Our claim is therefore not that CBTL-Transformers are universally preferable to AR models. Rather, we show that, when a transformer is used, CBTLs can make its internal representations more interpretable. The broader question of when transformers are preferable to simpler forecasting models is outside the scope of this work.
11. This is not concept leakage as commonly discussed in CBM literature, which means that other information beyond the pre-defined concepts passes through the bottleneck. Rather, this observation hints at the opposite: the same information from the bottleneck is *also* present in other model components. We hypothesize that it helps the model to have similar information in earlier layers to obtain a high CKA score in the bottleneck, and it does not hurt the forecasting performance enough to remove it in later layers.

Minor issues

- We acknowledged and added the references [2-7] ([8] was already included)
- Changed ‘concept per head’ to ‘heads per concept’ (page 6).
- When multiple heads share one concept, either (1) each head is included independently in the loss function with the shared concept, or (2) the representations of each head are concatenated for calculating CKA. We did not perform such experiments.

Presentation issues/ typos

- We fixed typos in the Fig. 2 legend, Fig. 10 caption, on page 8 and on page 11.
- We changed the location of tables and figures (incl. Fig. 8) to be closer to their in-text mention.

Dear Stefano Colamonaco,

We appreciate the time and effort you put into the detailed review. We considered all comments and changed the paper as follows:

Framing

We agree that the claim [CBMs are not usually considered neuro-symbolic] is too strong. We changed the background section. We see this paper as a contribution to concept-based interpretability for forecasting, taking inspiration from mechanistic interpretability. In fact, CBMs provide the opportunity for symbolic concept reasoning, e.g., by building syntactic rule structures [1], learning logic rules [2], or learning causal relations [3]. We adapted the introduction with the corrected framing.

Structure

We restructured the Background and Related Work section as suggested. The CKA definition and transformer formalism are moved to the Method section. The remainder has been reorganised thematically, and more literature has been brought to representational similarity measures used as training signals. We use the suggested formulation of “concurrent and subsequent work” and explicitly state what this manuscript adds relative to the earlier preprint.

What interpretability means here

Please note that our activation patching experiment is already an evaluation of our method without using CKA. By changing the hidden activations of the designated model component, we are directly testing its downstream effect.

Additionally, we performed the suggested probing experiments. Note that the set-up for probing is not trivial, because the concept labels are vectors (matching the input time series). To train probes on classification or regression, we opt for simpler tasks to test whether the designated model components contain some of the information. In particular, we train a probe on the final input hour-of-day and the final value that the trained AR model should predict from the input. These probes test whether the representations contain some of the input features and whether a correct representation of the trained AR model is made, respectively. The results are added to the experiments (on page 15).

The intervention

We changed the name of the experiment from ‘intervention’ to ‘activation patching’. We specify the dataset in contribution 3.

The bottleneck is not a bottleneck

We added concept leakage as a limitation in the discussion section on page 19, referring to Mahinpei et al. 2021.

Figures

We changed Figs. 1 and 7a as suggested.

Results

1. We repeated our experiments on five seeds and verified the table.
It is correct that the “No bottleneck” column differs between Tables 3 and 4 (now Table 5 and 6). We trained separate baseline models to match the architecture of the two bottleneck configurations. In Table 5, the bottleneck model has 3 components, resulting in 3 attention heads ($c=3$, with free component). In Table 6, the bottleneck has only 2 concepts, and therefore only 2 attention heads ($c=2$, without free component). We clarified this in the caption.
2. The Wu et al. column is indeed quoted from the original paper. The “no bottleneck” column is the baseline Autoformer model without CBTL, which has similar dimensions to the bottleneck models. We clarified this in the caption of Table 2.
3. We added a discussion on the difference in results for the Att and FF CBTL activation patching experiment. We added an experiment on the vanilla Transformer, where there is no difference in effectiveness between the Att and FF CBTL. We suspect that the architectural changes in the Autoformer attention heads (using auto-correlation) are responsible for the difference in results.

Underspecified details

We clarified the main text (equation 1, the AR and hour-of-day descriptions in the method section, caption of Table 1, implementation details). In short:

- The features for CKA are centered.
- The hour-of-day feature is a scalar, but CKA permits different feature dimensions.
- For AR surrogate training, we use the same input length as the original Autoformer setting (36 for Illness and 96 for the others).
- We use the stacked values of the input window weighted by its coefficients as AR concept representations.
- We will release the code upon publication.

[1] Barbiero, Pietro, et al. "Interpretable neural-symbolic concept reasoning." International Conference on Machine Learning. PMLR, 2023.

[2] Debot, David, et al. "Interpretable concept-based memory reasoning." Advances in Neural Information Processing Systems 37 (2024): 19254-19287.

[3] Moreira, Ricardo, et al. "Diconstruct: Causal concept-based explanations through black-box distillation." *Proceedings of Machine Learning Research* vol TBD:1–29, 2024.