
Belief State Representations in Transformer Models on Nonergodic Data

Junfeng Feng
University of
Aberdeen/South China
Normal University

Wanjie Zhong
University of
Aberdeen/South China
Normal University

Doroteya Stoyanova
University of Aberdeen

Lennart Finke
University of Göttingen
/ ETH Zürich

Organized by
Adam Shai, Paul Riechers, PIBBSS

Abstract

We extend research that finds representations of belief spaces in the activations of small transformer models, by discovering that the phenomenon also occurs when the training data stems from Hidden Markov Models whose hidden states do not communicate at all. Our results suggest that Bayesian updating and internal belief state representation also occur when they are not necessary to perform well in the prediction task, providing tentative evidence that large transformers keep a representation of their external world as well.

Keywords: *Transformer Models, Non-Ergodic Data, Mixed State Representations, Bayesian Updating, Sequential Data Modeling, Hidden Markov Models, Computational Mechanics.*

1. Introduction

Transformer models have demonstrated remarkable prowess in modelling sequential data across various domains. Recent studies have demonstrated that toy transformers have patterns in their activations that suggest that they keep an internal representation of a process producing the training data and perform Bayesian updating on it (Shai, 2024), producing geometric structures as studied in the field of computational mechanics (Shalizi, 2008). Our project discusses a similar setting, where a Hidden Markov Model (HMM) that produces the data is

nonergodic, meaning that not all hidden states can be visited from each other with positive probability.

The study focuses on two main questions: **(1) Can we replicate linear representations of belief spaces in the activations of transformers for non-ergodic processes?** **(2) In certain nonergodic processes, do models fail to represent the whole belief space but successfully represent subspaces?**

2. Methods

To investigate hypotheses, we trained Transformers on three processes below:

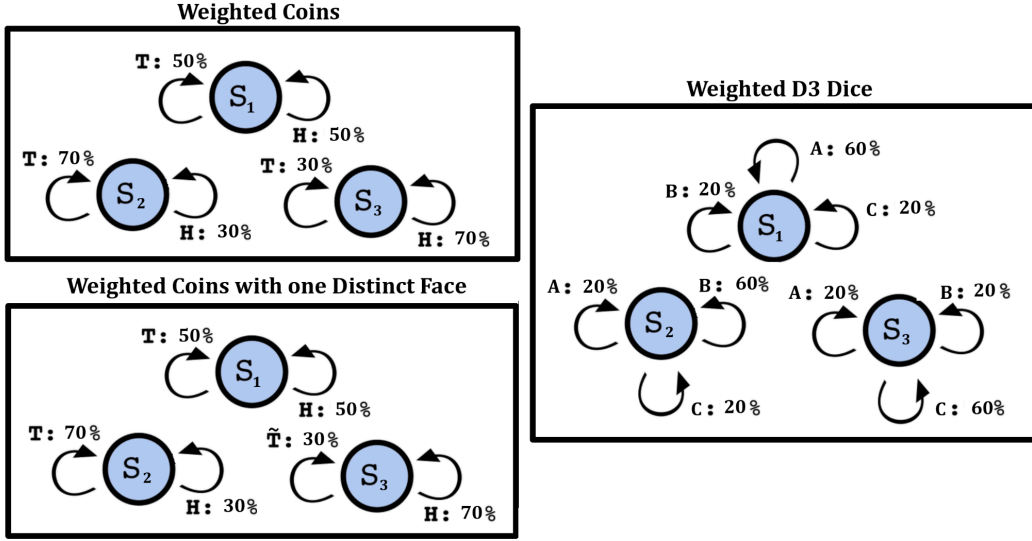


Figure 1 - HMMs for Weighted Coins, Weighted Coins with one Distinct Face and Weight D3 Dice

1. Weighted Coins. The hidden states are three indistinguishable coins with different probabilities. One is fair, the others are weighted to different tokens.
2. Weighted Coins with one Distinct Face. Same as above, but with one biased side emitting a different token, as if marking one side of the coin.
3. Weighted D3 Dice. Same as the coins, but with three outcomes, each weighted towards a different token. Like Mess3 in (Shai, 2024), but without transitions between the hidden states.

The transformer architecture and hyperparameters were kept consistent across all training runs to ensure fair comparisons. A context window of 6 was used for the first two processes, and for the third, a context window of 8.

To analyze the learned representations, we follow (Shai, 2024) and calculate the mixed state representations (MSPs) as dictated by Bayesian updating with complete knowledge of the HMM. We regress these belief states to the activation of the residual stream of the transformer given a minimal token sequence that leads to the belief state. The code for constructing the HMMs, generating the training data, training the transformer models, and analyzing the MSPs is available in our GitHub repository: https://github.com/doro041/CompMechanic_Transformer.

3. Results

3.1) We consider the Weighted D3 Dice process and answer the question (1) positively: Yes, it appears that even in this highly nonergodic case where none of the hidden states communicates, linear regression can recover a structure similar to the MSP belief space (Figure 2). We find this rather surprising because the optimal strategy (which should be easily encoded in the weights) is simply to count which token came up most and predict that. What benefit is there to tracking these probabilities? This could suggest transformers not only happen to do Bayesian updating, but they also prefer using it over simpler algorithms.

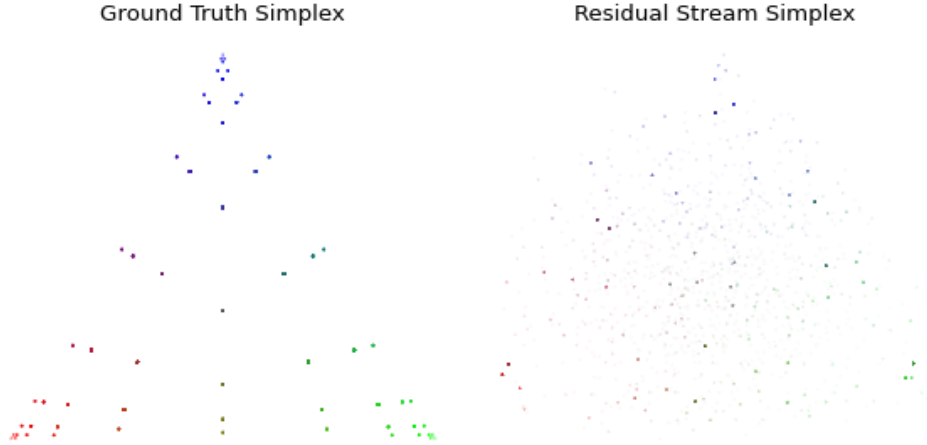


Figure 2 – Weighted Coins: Belief states by Bayes’ rule vs. transformer activations

3.2) The Weighted Coins process answers the question (2), again in the affirmative. We constructed the process to produce an interesting belief space but anticipated that this would not be learned, as shown in Figure 3. The reason might be that the information that the hidden state is “the fair coin” cannot be used by the transformer to minimise its loss; it can only predict heads or tails, and both are equally likely. However, there *is* a use for the information which biased coin the hidden state is in, so we find a linear subspace representing this. In particular, notice how the residual stream activations look like a projection on the simplex edge connecting the two biased coins.

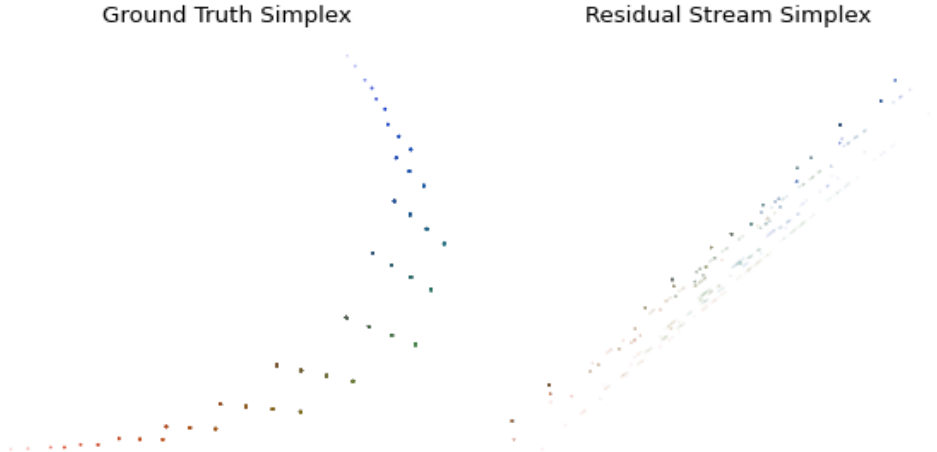


Figure 3 – Weighted D3 Dice: As in Figure 2

3.3) The Weighted Coins with one Distinct Face process illustrate partial success in representing a belief space, as depicted in Figure 4. We do not understand precisely why they are *where* they are, but it appears the model used two one-dimensional linear subspaces as its beliefs, which approximate the sparser, nonlinear belief space that complete information would dictate.

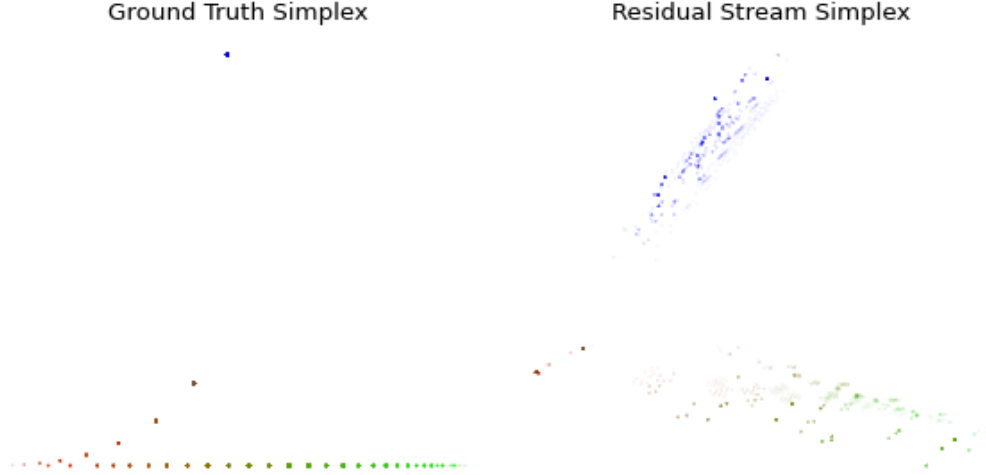


Figure 4 – Weighted Coins with one Distinct Face: As in Figure 2

4. Discussion and Conclusion

The most pressing question around the application of computational mechanics to transformers is whether its results are also applicable to numbers of parameters several orders of magnitudes larger. We hope to have produced some clues that even with processes such as a natural language corpus with many disjoint parts (such as different languages, topics, styles, etc.), internal belief state representations can occur, including when keeping one is not efficient. As a next step, we would be interested in exploring the mathematics phenomena 3.3), particularly, can we predict that the belief state will manifest as two distinct linear subspaces rather than a more complex structure? Additionally, given that all processes under consideration lack transient states, how will these processes be represented within this framework (Oden, 2003) ?

We thank the organisers for the opportunity to learn more about computational mechanics and perhaps even contribute a little to AI safety.

5. References

- [1] Shai, A. (2024) *Transformers represent belief state geometry in their residual stream*, *LessWrong*. Available at: <https://www.lesswrong.com/posts/gTZ2SxesbHekJ3CkF/transformers-represent-belief-state-geometry-in-their> (Accessed: 03 June 2024).
- [2] Shalizi, C.R. and Crutchfield, J.P. (2008) *Computational mechanics: Pattern and prediction, structure and simplicity - journal of statistical physics*, *SpringerLink*. Available at: <https://link.springer.com/article/10.1023/A:1010388907793> (Accessed: 03 June 2024).
- [3] Oden, J.T., Belytschko, T., Babuska, I. and Hughes, T., 2003. Research directions in computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, 192(7-8), pp.913-922. Available at: <https://www.sciencedirect.com/science/article/pii/S0045782502006163> (Accessed: 03 June 2024).