Exploring Hierarchical Structure Representation in Transformer Models through Computational Mechanics[1]

Aayush Kucheria
Aalto University, Finland

Olli Järviniemi
Independent Researcher, Finland

Udayanto Dwi Atmojo
Aalto University, Finland

Konsta Tiilikainen
Aalto University, Finland

Organized by
Aalto EA, Adam Shai, Paul Riechers, PIBBSS

Abstract

Interpretability, a key property of trustworthy AI, allows one to understand the internal workings of AI systems. Previous work in interpretability within the explainable AI body of knowledge hasn't focused on a deep understanding of AI systems, leading to a "black-box" problem. Mechanistic interpretability attempts to address this gap, and there have also been recent approaches that utilize computational mechanics to more formally engage with the internal phenomena. However, this domain is still in its infancy, where, for example, the data generating processes considered, are relatively simple and it remains unknown whether the approaches can scale as the (data-generating) process gets more complex.

This work attempts to explore how an approach based on computational mechanics can cope when a more complex hierarchical generative process is involved, i.e, a process that

[1] Research conducted at the Computational Mechanics For AI Safety Hackathon, 2024 (https://www.apartresearch.com/event/compmech)

comprises Hidden Markov Models (HMMs) whose transition probabilities *change* over time.

We find that small transformer models are capable of modeling such changes in an HMM. However, our preliminary investigations did not find geometrically represented probabilities for different hypotheses.

Future work could include performing more systematic exploration of the types of hierarchical processes studied in this work, and to more confidently test whether the representations generated by the computational mechanics framework are able to model more complex phenomena.

Keywords: Computational Mechanics, Transformers, Interpretability, AI Alignment.

1. Introduction

Artificial intelligence (AI) is a rapidly advancing domain, with a wide variety of use cases and application areas. However, its rapid growth creates various concerns among stakeholders, from the general public to decision makers, on its potential negative impacts to the society. Following this concern, there have been various research, development, and innovation (RDI) initiatives to approach the so-called trustworthy AI.

The policy report of the EU Joint Research Centre [1] views "trustworthy" as a "comprehensive framework" which accounts for various principles, requirements and criteria, meant for creating AI systems which are "human-centric and rest on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom". It is expected that when trustworthiness is considered, it would maximise the benefits of AI systems while at the same time preventing and minimising their risks [2].

Recent advances in neural network-based AI show a trend towards increasingly more complex architectures with the hope that they can solve more complex problems. In the last several years, transformer-based models such as large language models (LLMs) have made an emergence. While these models have shown massive potential in various areas, they have been considered as "black-box" models where it is challenging to assess their trustworthiness.

One key property within trustworthy AI is interpretability. In the literature found within the explainable AI (XAI) field, interpretability and explainability have been used interchangeably. However, interpretability has been increasingly considered as a prerequisite for explainability [3] [4]. Some popular definitions of interpretability

are highlighted in [5] and [6], such as "the ability to explain or to present in understandable terms to a human", albeit lacking any formal mathematical representations. XAI practitioners hypothesize that increased degree of interpretability will correspond to easier identification of cause-effect relationship between input and output of a model [7].

There have been various proposals of methods to assess interpretability by the explainable AI (XAI) community. E.g., the work done in [7] analyzes various methods and provides a taxonomy of interpretability methods. However, the methods proposed by the explainable AI community have been criticized as they are unable to show how trained AI models work overall [8]. Mechanistic interpretability is a new emerging field which is investigating ways that highlight how trained AI models are functioning by reverse engineering circuits in trained neural networks into forms understandable by humans. [9] makes an attempt to review the existing body of knowledge within mechanistic interpretability, but it offers limited attention to the development within another emerging research direction utilizing computational mechanics.

Recent developments in computational mechanics seem to offer exciting and interesting possibilities on assessing interpretability, e.g., how one attempts to understand the internals of "black-box" AI models such as transformers. It provides formalizations for modeling the computational structures that emerge within LLMs. Using these formalization tools, [10] attempts to make specific theoretical predictions about the geometry of residual stream activations in LLMs and test them empirically. They consider the training data generated by a simple Hidden Markov Model (HMM) called the Mess3 process. This process has 3 states, and different probabilities that define the transitions between them. However, this data generating process is extremely simple in comparison to the generative processes behind the training data used in real-world LLMs. This observation leads to the following research question (RQ):

RQ: Is the approach discussed in [10] able to cope with more complex hierarchical data-generating processes?

In this work, we aim to build on [10] by considering a more complex (hierarchical) data generating process. Our goal is to investigate if the approach proposed in [10] can be used for transformer models that model "deep" latents, such as latents with hierarchical structure.

We hope that this work can contribute to the existing body of knowledge within computational mechanics, and the broader field of interpretability. This report is organized as follows. Section 2 describes the methodology and experiments considered in this work. Section 3 presents the results of the experiments we do, followed by discussions, and conclusions and potential future works in Section 4.

2. Methods

Data generating process

Our data generating process is a Hidden Markov Model whose transition probabilities *change* over time. The HMM generates a binary sequence of data for a transformer to predict. We randomly choose one of two ways of changing transition probabilities for each input sequence. This incentivizes the model to not only reason about the hidden states, but also about the way the transition probabilities are changing over time.

The HMM has two states A and B. The transition probabilities between these states change linearly as the output characters 0 and 1 are sampled from the model. More formally, the probabilities are determined by two parameters t and *sign*, where t is the length of the context history so far, and *sign* is uniformly sampled from {-1, 1}. Note that t changes during inference time, while *sign* is fixed latent variable the transformer can perform probabilistic inference on. The data generating process we used is illustrated in Figure 1.

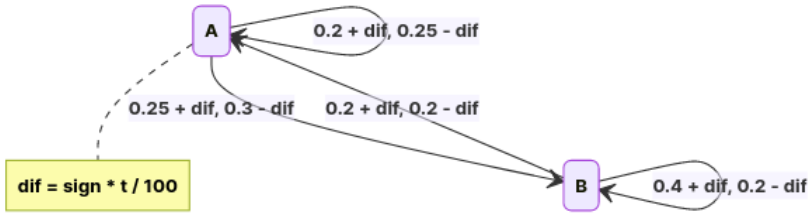

Figure 1: Visualization of evolving transition probabilities in a Hidden Markov Model (HMM) with two states, A and B. Each transition arrow shows two probabilities: the first for the transition and outputting a 0, the second for the transition and outputting a 1.

These probabilities change linearly based on the sequence length (t) and a uniformly sampled sign parameter (sign), facilitating complex probabilistic inferences by the transformer.

Thus, for the transformer to obtain optimal predictive performance, it needs to perform probabilistic inference not only on the hidden states A and B, but also on the *sign* parameter. The data generating process naturally suggests considering four hypotheses corresponding to the two hidden states and two possible values of *sign*.

Training and evaluation

We trained the model on 66k sequences of length 19 for 300 epochs sampled by the process described above. The model was a transformer model with 4 layers and residual stream dimension 16.

To measure the model's performance, we computed cross-entropy loss on a validation set. We compared the achieved loss to two points of reference: the "*minimum loss*" achieved by a predictor performing ideal Bayesian updates on the current hidden state and the value of *sign*, and the "*trivial loss*" achieved by a predictor assuming the transition probabilities do not change during inference (i.e. assuming $sign = 0$).

We also looked for geometric representations of the hypotheses in the model's residual stream. We took linear projections at the last token place in the last layer, aiming to find a projection that matches with the probabilities assigned to different hypotheses by the ideal Bayesian predictor.

We implemented our code in a Google Colab notebook, which can be found in this url address[2].

3. Results and Discussions

Performance

We evaluated the model on 33k sequences sampled by the same process as the training data. We found that the model obtains essentially the minimum loss, which is considerably better than trivial loss. Thus, the model is performing close to perfect probabilistic inference, and is *not* merely modeling the data as a Hidden Markov Model.

```
Test Loss: 99.36067393415189 percent of minimum, 89.05841199925688 percent of baseline,
```

(The model appears to be slightly better than the optimal predictor; this is likely due to random fluctuation and/or an off-by-one error in token indices when computing losses. In any case, we are confident that the model attains better than trivial loss.)

Internal representations

However, we could not find linear projections that accurately determine probabilities assigned by ideal predictors. This is largely due to time constraints and implementation-related issues; we couldn't perform good enough checks to be confident in our methods, and faced difficulties when visualizing our findings.

It is also possible that the model is not representing such information in the format we expected; for example, the model could in principle *separately* represent hypotheses corresponding to the value of the hidden state and hypotheses corresponding to the *sign* parameter ("P(hidden state = A) = p, P(sign = 1) =

[2] https://colab.research.google.com/drive/16BdwjQTi7uSL-OWleRKIqx3oHDnVxEfz?usp=sharing

q"), rather than representing the combined hypotheses ("P(hidden state = A, sign = 1) = r").

Finally, we only tested projecting the last layer of the residual stream (and only at the last token position), but it is possible that the representations for such hypotheses are distributed over several layers.

4. Conclusions and Potential Future Works

Computational Mechanics can act as a powerful lens to probe AI models with. It offers a formalized way to think about the dynamics of learning the underlying structures of the data, and provides theoretical predictions to test against.

In this work, we built upon some early work in this field. We defined a more complex data generative process, with hierarchical latents, and sampled data from it to test. The results of our work appeared to provide some answers, or at least some hints towards answers, to our RQ, however, more systematic explorations and experiments are needed.

For future work, we'd like to see more systematic exploration of the types of hierarchical latents studied in this work, and to more confidently verify that 1) transformers can model deep (hierarchical) latents and 2) whether their beliefs about such latents are linearly represented in the residual stream. More broadly, constructing more nested data generating processes allows for studying the depth of neural networks' models of the world.

5. References

[1] D. Fernández Llorca and E. Gómez, "Trustworthy autonomous vehicles," *Publ. Off. Eur. Union Luxemb. EUR*, vol. 30942, 2021.

[2] N. A. Smuha, "The EU approach to ethics guidelines for trustworthy artificial intelligence," *Comput. Law Rev. Int.*, vol. 20, no. 4, pp. 97–106, 2019.

[3] F. Doshi-Velez *et al.*, "Accountability of AI Under the Law: The Role of Explanation." arXiv, Dec. 20, 2019. doi: 10.48550/arXiv.1711.01134.

[4] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *Int. J. Comput. Vis.*, vol. 130, no. 10, pp. 2425–2452, 2022.

[5] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *ArXiv Prepr. ArXiv170208608*, 2017.

[6] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.

[7] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, 2021, doi: 10.3390/e23010018.

[8] L. Kästner and B. Crook, "Explaining AI Through Mechanistic

Interpretability," 2023.
[9] L. Bereska and E. Gavves, "Mechanistic Interpretability for AI Safety--A Review," *ArXiv Prepr. ArXiv240414082*, 2024.
[10] A. S. Shai, S. E. Marzen, L. Teixeira, A. G. Oldenziel, and P. M. Riechers, "Transformers represent belief state geometry in their residual stream." arXiv, May 24, 2024. doi: 10.48550/arXiv.2405.15943.