How Do Transformers “Do” Physics? Investigating the Simple Harmonic Oscillator

Anonymous Authors[1]

Abstract

How do transformers model physics? We take a step to demystify this larger puzzle by investigating how transformers model the simple harmonic oscillator (SHO), $\ddot{x} + 2\gamma\dot{x} + \omega_0^2 x = 0$, one of the most fundamental systems in physics. Our goal is to identify the methods transformers use to model the SHO, and to do so we hypothesize and evaluate possible methods by analyzing the encoding of these methods' intermediates. We develop two correlational and two causal criteria for the use of a method within the simple testbed of linear regression, where our method is $y = wx$ and our intermediate is w. Armed with these four criteria, we determine that transformers use known numerical methods to model trajectories of the simple harmonic oscillator, specifically the matrix exponential method. Our analysis framework can conveniently extend to high-dimensional linear systems and nonlinear systems, which we hope will help reveal the “world model” hidden in transformers.

1. Introduction

Transformers are state of the art models on a range of tasks (1; 2; 3; 4), but our understanding of how these models represent the world is limited. Recent work in mechanistic interpretability (5; 6; 7; 8; 9; 10; 11; 12) has shed light on how transformers represent mathematical tasks like modular addition (13; 14; 7), yet little work has been done to understand how transformers model physics. This question is crucial, as for transformers to build any sort of “world model,” they must have a grasp of the physical laws that govern the world (15)[1].

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Transformers with better “world models” are also at greater risk of misuse.

Our key research question is: How do transformers model physics? This question is intimidating, since even humans have many different ways of modeling the same underlying physics (16). In the spirit of hypothesis testing, we reformulate the question as: given a known modeling method g, does the transformer learn g? If a transformer leverages g, its hidden states must encode information about important intermediate quantities in g. We focus our study on the simple harmonic oscillator $\ddot{x}+2\gamma\dot{x}+\omega_0^2 x = 0$, where γ and ω_0 are the damping and frequency of the system respectively. Given the trajectory points $\{(x_0, v_0), (x_1, v_1), \ldots, (x_n, v_n)\}$ at discrete times $\{t_0, t_1, \ldots, t_n\}$, we task a transformer with predicting (x_{n+1}, v_{n+1}) at time t_{n+1}, as shown in Fig. 1. In this setting, g could be a numerical simulation the transformer runs after inferring γ, ω_0 from past data points. We would then expect some form of γ and ω_0 to be intermediates encoded in the transformer. How can we show that intermediates and the method g are being used?

We develop criteria to demonstrate the transformer is using g by studying intermediates in a simpler setting: in-context linear regression, $y = wx$. As correlational evidence for the model's internal use of w, we find that the intermediate w can be encoded linearly, nonlinearly, or not at all. We also link the performance of models to their encoding of w and use it as an explanation for in-context learning. We generate causal evidence for the use of w by analyzing how much of the hidden states' variance w explains and linearly intervening on the network to predictably change its behavior.

We use these developed criteria of intermediates to study how transformers model the simple harmonic oscillator (SHO), a fundamental model in physics. We generate multiple hypotheses for the method(s) transformers use to model the trajectories of SHOs, and use our criteria from linear regression to show correlational and causal evidence that transformers employ known numerical methods, specifically the matrix exponential, to model trajectories of SHOs. Although our analysis is constrained to the SHO in this paper, our framework naturally extends to some high-dimensional linear and nonlinear systems.

Organization In Section 2 we define and investigate intermediates in the setting of linear regression and use this to develop criteria for transformers' use of a method g. In Section 3 we hypothesize that transformers use numerical methods to model the SHO, and use our criteria of intermediates to provide causal and correlational evidence for transformers' use of the matrix exponential. Due to limited space, related work is deferred to Appendix A.

2. Developing Criteria for Intermediates with Linear Regression

Our main goal is to determine which methods transformers use to model the simple harmonic oscillator. We would like to do this by generating criteria based on the encoding of relevant intermediates. In this section, we develop our criteria of intermediates in a simpler setting: linear regression. Notably, linear regression is identical to predicting the acceleration from the position of an undamped harmonic oscillator ($\gamma = 0$), making this setup physically relevant.

Setup In our linear regression setup, we generate $\boldsymbol{X}$ and $\boldsymbol{w}$ between $[-0.75, 0.75]$, where $\boldsymbol{X}$ has size $(5000, 65)$ and $\boldsymbol{w}$ has size $(5000,)$. We generate $\boldsymbol{Y} = \boldsymbol{w}\boldsymbol{X}$, and train transformers to predict y_{n+1} given $\{x_1, y_1, ..., x_n, y_n, x_{n+1}\}$.

Since in-context linear regression is well studied for transformers (17; 18), we use this simple setting to ask and answer fundamental questions about intermediates, namely:

- **What** is an intermediate?
- **How** can intermediates be encoded and how can we robustly probe for them?
- **When**, or under what circumstances, are intermediates encoded?

All of these questions develop an understanding of intermediates that builds up to the **Key Question: How can we use intermediates to demonstrate that a transformer is actually using a method in its computations?** By answering this question for linear regression, we generate four correlational and causal criteria to demonstrate a transformer is using a method in its computations, which we can then apply to understand the simple harmonic oscillator, as shown in Fig. 1.

2.1. What is an intermediate?

We define an intermediate as a quantity that a transformer uses during computation, but is not a direct input/output to/of the transformer. More formally, if the input to the transformer is $\boldsymbol{X}$ and its output is $\boldsymbol{Y}$, we can model the transformer's computation as $\boldsymbol{Y} = g(\boldsymbol{X}, I)$, where g is the method used and I is the intermediate of that method. For example, if we want to determine if the transformer is computing the linear regression task using $\boldsymbol{Y} = \boldsymbol{w}\boldsymbol{X}$, then $I = \boldsymbol{w}$, $g(\boldsymbol{X}, I) = g(\boldsymbol{X}, \boldsymbol{w}) = \boldsymbol{w}\boldsymbol{X}$.

2.2. How can intermediates be encoded and how can we robustly probe for them?

We want to understand what form of the intermediate, $f(I)$, is encoded in the network's hidden states. For example, while it may be obvious to humans to compute $y = wx$, perhaps transformers prefer $\exp(\log(w) + \log(x))$ or $\sqrt{w^2x^2}$. We want to develop a robust probing methodology that captures these diverse possibilities. We identify three ways an intermediate I can be represented: linearly encoded, nonlinearly encoded, and not encoded at all. We use HS to mean hidden state.

Linearly encoded We say I is linearly encoded in a hidden state HS if there is a linear network that takes $I = \text{Linear}(HS)$. We determine the strength of the linear encoding by evaluating how much of the variance in I can be explained by HS, i.e. the R^2 of the probe.

Nonlinearly encoded To probe for an arbitrary $f(I)$, we define a novel **Taylor probe**, which finds coefficients a_i such that $f(I) = a_1 I + a_2 I^2 + ... + a_n I^n$, and $f(I) = \text{Linear}(HS)$. To actually implement this probing style, we use Canonical Correlation Analysis probes, which given some multivariate data X and Y, finds directions within X and Y that are maximally correlated (19). Here, $X = [I, I^2, I^3, ..., I^n]$, and $Y = HS$. If I is of bounded magnitude and n is sufficiently large, we are able to probe the transformer for any function $f(I)$. In practice, we use $n \leq 5$.

Not encoded If I fails to be linearly or nonlinearly encoded, we say that it is not encoded within the network. For example, there are at least two ways to predict y_2 from $\{x_1, y_1, x_2\}$ such that $y_2 = \frac{y_1}{x_1} x_2$. (1) $w = y_1/x_1$ is encoded, and $y_2 = wx_2$. (2) $w' = x_2/x_1$ is encoded (so $w = y_1/x_1$ is not encoded), and $y_2 = w' y_1$. Thus, it is not guaranteed that w is encoded.

2.3. When, or under what circumstances, are intermediates encoded?

We want to apply our probing techniques to better understand what type of models generate intermediates. Under the described setting of linear regression, we train transformers of size $L = [1, 2, 3, 4, 5]$ and $H = [2, 4, 8, 16, 32]$ [2], where L is the number of layers and H is the hidden size of the transformer. We find that these models generalize to

[2] All transformers trained in this study use one attention head and no LayerNorm to aid interpretability, and are trained on a NVIDIA Volta GPU with the hyperparameters $\text{epochs} = 20000, \text{lr} = 10^{-3}, \text{batchsize} = 64$ using the Adam optimizer (20).

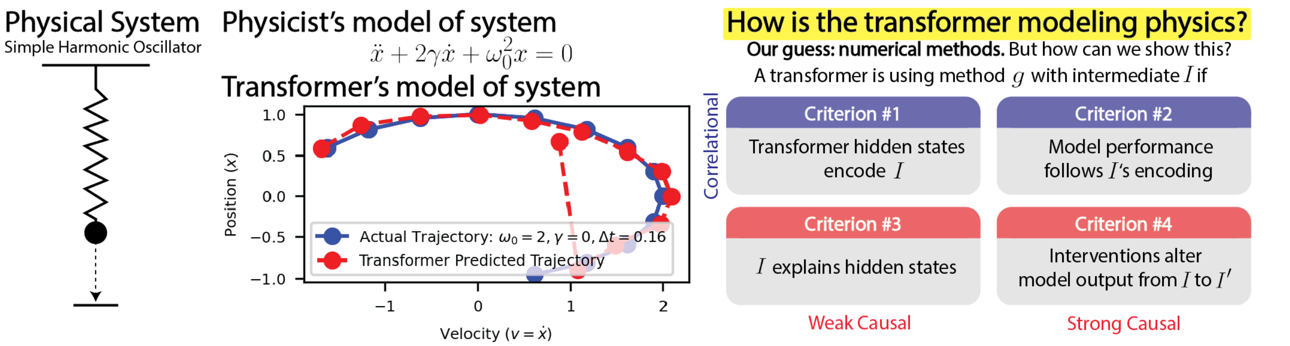

Figure 1. We aim to understand how transformers model physics through the study of meaningful intermediates. We train transformers to model simple harmonic oscillator (SHO) trajectories and use our developed criteria of intermediates to show that transformers use known numerical methods to model the SHO.

out-of-distribution test data ($0.75 \leq |\boldsymbol{w}| \leq 1$) in Appendix Fig. 10, but we focus on investigating intermediates on in-distribution training data.

Larger models have stronger encodings of intermediates We find that smaller models often don't have $\boldsymbol{w}$ encoded, while larger models encode $\boldsymbol{w}$ linearly, as evidenced by Fig. 2. We formalize this further by defining $\max(\bar{R^2})$ as the maximum value taken over depth positions of the mean R^2 of $\boldsymbol{w}$ probes taken over context length. In Appendix Fig. 12, we observe a clear phase transition in encoding across model size, and also find that $\max(\bar{R^2})$ does not significantly improve if we extend the degree of the Taylor probes to $n > 2$. Thus, in the case of linear regression, we find that models represent $\boldsymbol{w}$ linearly, quadratically, or not at all.

We attribute the stronger encoding of $\boldsymbol{w}$ in larger models to the "lottery ticket hypothesis" - larger models have more "lottery tickets" in their increased capacity to find a "winning" representation of $\boldsymbol{w}$ (21; 22). Interestingly, the intuitive understanding that larger models have $\boldsymbol{w}$ better encoded leads us to the counterintuitive conclusion that larger models are actually *more* interpretable for our purposes.

Encoding quality is tied to model performance In Appendix Fig. 13, we find that better performing models generally have stronger encodings of $\boldsymbol{w}$. In Fig. 3, we also find that the improvements in model prediction as a function of context length, or in-context learning, are correlated to improvements in $\boldsymbol{w}$'s encoding, which we would expect if our models were using $\boldsymbol{w}$ in their computations.

2.4. Key Question: How can we use intermediates to demonstrate that a transformer is actually using a method in its computations?

So far we have discovered that models encode $\boldsymbol{w}$, either linearly or nonlinearly, and found relationships between model size, performance, and encoding strength. But how can we

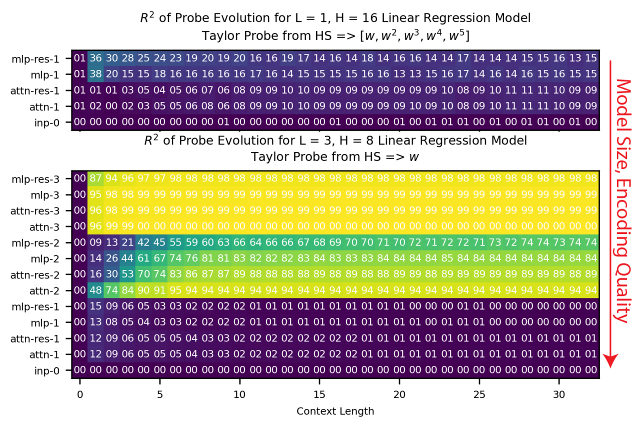

Figure 2. We plot the R^2 of Taylor probes for the intermediate $\boldsymbol{w}$ within models trained on the task $\boldsymbol{Y} = \boldsymbol{w}\boldsymbol{X}$. We see that larger models often have $\boldsymbol{w}$ encoded linearly, while smaller models do not have $\boldsymbol{w}$ encoded, even for high Taylor probe degree.

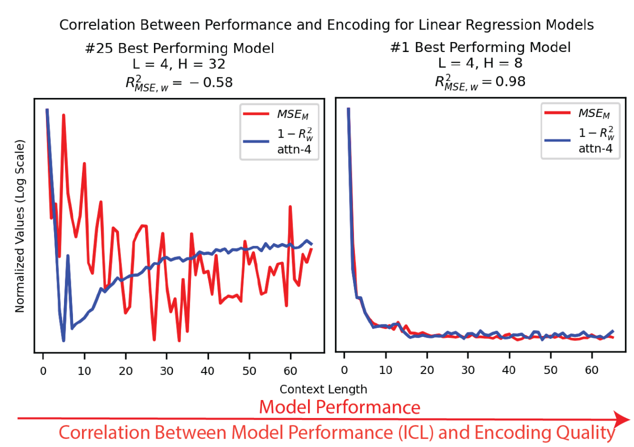

Figure 3. We find that the ability of the best performing models to in-context learn is highly correlated with their encoding of $\boldsymbol{w}$ ($R^2(MSE, w)$). We plot normalized values for the error of the encoding ($1 - R^2_w$) in red and the mean squared error of the model (MSE_M) in blue.

ensure that the model is actually using $\boldsymbol{w}$ in its computations and the encoding of $\boldsymbol{w}$ is not just a meaningless byproduct (23)?

Reverse Probing To ensure that $\boldsymbol{w}$ is not encoded in some insignificant part of the residual stream, we set up probes going from $[w, w^2] \to HS$, as opposed to $HS \to f(w)$. In Fig. 4, we often find that $\boldsymbol{w}$ can explain large amounts of variance in model hidden states, implying that these hidden states are dedicated to representing $\boldsymbol{w}$. We take this as weak causal evidence that $\boldsymbol{w}$ is being used by the model - otherwise, it is unclear why a part of the model would be dedicated to storing $\boldsymbol{w}$.

Intervening We can also use reverse probes to intervene on the models' hidden states and predictably change their output from $\boldsymbol{w} \to \boldsymbol{w'}$. In Fig. 4 we attempt to make $\boldsymbol{w'} = 0.5$ for all series, and then measure the observed $\hat{\boldsymbol{w}}$ from the models' outputs ($\hat{w} = \hat{y}_n / x_n$). For 4 out of 25 models the intervention worked, providing strong causal evidence that the model uses its internal representation of $\boldsymbol{w}$ in computations. For models where we identified a quadratic representation of $\boldsymbol{w}$, we see that $\boldsymbol{w} = 0.5, -0.5$ are both represented in the observed intervention.

Putting it all together We can generalize our understanding of intermediates from linear regression to create criteria for a transformer's use of a method g in its computations.

Criteria for use of a method g with an associated, unique intermediate I

1. If a model uses a method g, its hidden states should encode I (shown in Fig. 2).
2. If a model uses a method g, model performance should improve if I is better represented (shown in Fig. 3).
3. If and only if the model uses g, we expect some hidden state's variance to be almost fully explained by I (shown in Fig. 4).
4. If and only if the model uses g, we can intervene on hidden states to change $I \to I'$ and predictably change the model output from $g(\boldsymbol{X}, I) \to g(\boldsymbol{X}, I')$ (shown in Fig. 4)

The first two criteria for a transformer's use of g are correlational and the last two are weak and strong causal. Using these criteria (summarized in Fig. 1), we can now investigate how transformers model more complex systems like the simple harmonic oscillator.

3. Investigating the Simple Harmonic Oscillator

We now apply our developed criteria of intermediates to investigate how transformers represent physics, specifically the methods they use to model the simple harmonic oscillator (SHO). The simple harmonic oscillator is ubiquitous in physics, used to describe phenomena as diverse as the swing of a pendulum, molecular vibrations, the behavior of AC circuits, and quantum states of trapped particles. Given a series of position and velocity data of a simple harmonic oscillator at a sequence of timesteps, we ask

1. Can a transformer successfully predict the position/velocity at the SHO's next timestep?
2. Can we determine what computational method the transformer is using in this prediction?

3.1. Mathematical and computational setup

The simple harmonic oscillator is governed by the linear ordinary differential equation (ODE)

$$\ddot{x} + 2\gamma\dot{x} + \omega_0^2 x = 0. \tag{1}$$

The two physical parameters of this equation are γ, the damping coefficient, and ω_0, the natural frequency of the system. An intuitive picture for the SHO is a mass on a spring that is pulled from its equilibrium position by some amount x_0 and let go, as visualized in Fig. 1. ω_0 is related to how fast the system oscillates, and γ is related to how soon the system decays to equilibrium from the internal resistance of the spring. We focus on studying how a transformer models the undamped harmonic oscillator, where $\gamma = 0$. Given some initial starting position (x_0), velocity (v_0), and timestep Δt, the time evolution of the undamped harmonic oscillator is

$$\begin{aligned} x_k &= x_0 \cos(k\omega_0\Delta t) + \frac{v_0}{\omega_0}\sin(k\omega_0\Delta t) \\ v_k &= v_0 \cos(k\omega_0\Delta t) - \omega_0 x_0 \sin(k\omega_0\Delta t), \end{aligned} \tag{2}$$

where $v = \frac{dx}{dt}$. We generate 5000 timeseries of 65 timesteps for various values of $\omega_0, \Delta t, x_0$, and v_0, described in Appendix D. Following the procedure for linear regression, we train transformers of size $L = [1, 2, 3, 4, 5]$ and $H = [2, 4, 8, 16, 32]$ to predict (x_{n+1}, v_{n+1}) given $\{(x_0, v_0), (x_1, v_1), ...(x_n, v_n)\}$. In Appendix Fig. 15, we see that our transformers are able to accurately predict the next timestep in the timeseries of out-of-distribution test data, and this prediction gets more accurate with context length (i.e. in-context learning). But how is the transformer modeling the simple harmonic oscillator internally?

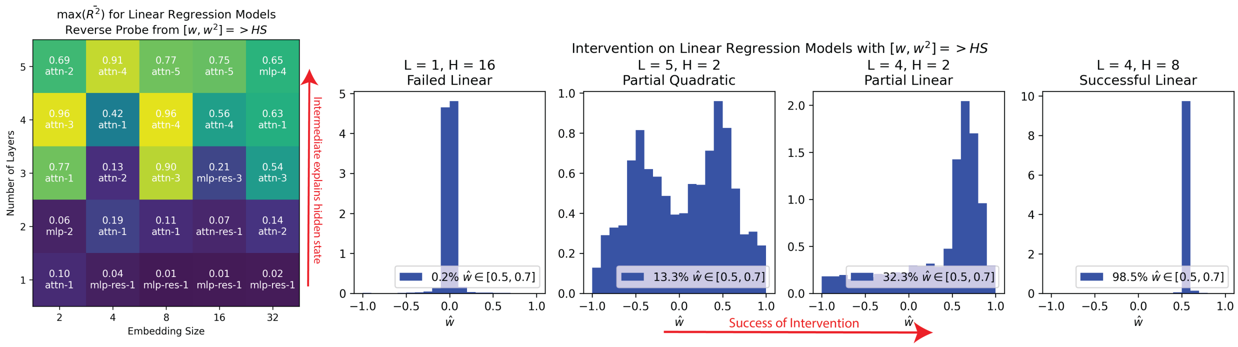

Figure 4. (Left) We plot $\max(\bar{R}^2)$ of the reverse probe from $[w, w^2] \to HS$ across all models, and find that the intermediate $\boldsymbol{w}$ can explain significant amounts of variance in model hidden states. (Right) We intervene using reverse probes to make all models output $\boldsymbol{w'} = 0.5$. This intervention can either fail (16/25), be partially successful nonlinearly (2/25) or linearly (3/25), or be successful (4/25).

3.2. What methods could the transformer use to model the simple harmonic oscillator?

Human physicists would model the simple harmonic oscillator with the analytical solution to Eq. 1, but it is unlikely that a transformer would do so. Transformers are numerical approximators which use statistical patterns in data to make predictions, and in that spirit, we hypothesize transformers use numerical methods to model SHOs. There is a rich literature on numerical methods that approximate solutions to linear ordinary differential equations (24; 25; 26), and we highlight three possible methods the transformer could be using in our theory hub. For notation, we note that Eq. 1 can be written as

$$\begin{bmatrix} \dot{x} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -2\gamma \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} x \\ v \end{bmatrix}. \tag{3}$$

Linear Multistep Method Our model could be using a linear multistep method, which uses values of derivatives from several previous timesteps to estimate the future timestep. We describe the kth order linear multistep method in Table 1 with coefficients α_j and β_j.

Taylor Expansion Method The model could also be using higher order derivatives from the previous timestep to predict the next timestep [3]. We describe the kth order Taylor expansion in Table 1.

Matrix Exponential Method While the two methods presented above are useful approximations for small Δt, the matrix exponential uses a 2×2 matrix to exactly transform the previous timestep to the next timestep. We describe it in Table 1. This method is the $\lim_{k\to\infty}$ of the Taylor expansion method.

In order to use the criteria described in Section 2 to figure out which method(s) our model is using, we need to define relevant intermediates for each method g. Similarly

Table 1. Numerical methods for modeling the simple harmonic oscillator and their intermediates.

Method	$\boldsymbol{g(X, I)}$	$\boldsymbol{I}$
Linear Multistep	$\begin{bmatrix} x_{n+1} \\ v_{n+1} \end{bmatrix} = \sum_{j=0}^{k} (\alpha_j + \beta_j \boldsymbol{A}\Delta t) \begin{bmatrix} x_{n-j} \\ v_{n-j} \end{bmatrix}$	$\boldsymbol{A}\Delta t$
Taylor Expansion	$\begin{bmatrix} x_{n+1} \\ v_{n+1} \end{bmatrix} = \sum_{j=0}^{k} \boldsymbol{A}^j \frac{\Delta t^j}{j!} \begin{bmatrix} x_n \\ v_n \end{bmatrix}$	$(\boldsymbol{A}\Delta t)^j$
Matrix Exponential	$\begin{bmatrix} x_{n+1} \\ v_{n+1} \end{bmatrix} = e^{\boldsymbol{A}\Delta t} \begin{bmatrix} x_n \\ v_n \end{bmatrix}$	$e^{\boldsymbol{A}\Delta t}$

to linear regression, the intermediates are the coefficients of the input, but are now 2×2 matrices and not a single value. We summarize our methods and intermediates in our theory hub in Table 1. Notably, these methods are viable for any homogeneous linear ordinary differential equation with constant coefficients, and potentially non linear differential equations as well (see Appendix C).

3.3. Evaluating methods for the undamped harmonic oscillator

We apply the four criteria established for linear regression (Fig. 1) to evaluate if transformers use the methods in Table 1. For the Taylor expansion intermediate, we use $j = 3$ to distinguish it from the linear multistep method, although our results are generally robust to $j \leq 5$ (Appendix Fig. 17). We summarize our evaluations across methods and criteria in Table 2.

Criterion 1: Is the intermediate encoded? In Fig. 5, we see that all three intermediates are well encoded in the model, with the matrix exponential method especially prominent. This provides initial correlational evidence that the models are learning numerical methods. The magnitude of encodings are generally smaller than the linear regression case, which we attribute to the increased difficulty of encoding 2×2 matrices compared to a single weight value

[3] This is equivalent to the nonlinear single step Runge-Kutta method for a homogenous linear ODE with constant coefficients

$\boldsymbol{w}$. Notably, we only probe for linear encodings given that $\boldsymbol{w}$ was most often encoded linearly in the linear regression case study.

Criterion 2: Is the intermediate encoding correlated with model performance? In Fig. 6, we see that for all three methods, better performing models generally have stronger encodings and worse performing models have weaker encodings. This correlation is strongest for the matrix exponential method. This provides more correlational evidence that our models are using the described methods.

Criterion 3: Can the intermediates explain the models' hidden states? In Fig. 7, we reverse probe from the intermediates to the models' hidden states, and find that all methods explain non trivial variance in model hidden states, while the matrix exponential method consistently explains the most variance by a sizable margin. This provides little weak causal evidence that the models are using the linear multistep and Taylor expansion methods, and stronger weak causal evidence that the models are using the matrix exponential method.

Criterion 4: Can we predictably intervene on the the model? *Criterion 4.1* To intervene on the model, we use the reverse probes from Fig. 7 to generate predicted hidden states from each intermediate. In Fig. 8, we then insert these hidden states back into the model and see if the model is still able to model the SHO. The matrix exponential method has the most successful interventions by an order of magnitude, and 18/25 of these intervened models perform better than guessing. This implies that the information the transformer uses to model the SHO is stored in the matrix exponential's intermediate.

Criterion 4.2 We can also vary $\Delta t \to \Delta t', \omega_0 \to \omega_0'$, regenerate intermediates and then hidden states, insert these modified hidden states into the model, and see if the model makes predictions as if it "believes" the input SHO data uses $\Delta t', \omega'$. In Fig. 9, we perform this intervention on Δt, but our results are robust to intervening on ω_0 as well (Appendix Fig. 18). Even for the model with the best reverse probe quality for the linear multistep/Taylor expansion intermediates ($L = 4, H = 4$), intervening with the matrix exponential method is most successful. Combined with our previous intervention (4.1), we now have strong causal evidence for the matrix exponential method.

The transformer likely uses the matrix exponential to model the undamped harmonic oscillator We have correlational evidence that the model is using all three methods in our theory hub, with little causal evidence for the linear multistep and Taylor expansion methods, and strong causal evidence for the matrix exponential method. We suspect the model is *only* using the matrix exponential method in its computations, and the evidence we have for the other two methods is a byproduct of the use of the matrix exponential. In Appendix Fig. 19, we give correlational evidence for this claim by generating synthetic hidden states from $e^{\boldsymbol{A}\Delta t}$ and showing that in this synthetic setup, we retrieve values for criterions 1, 3 for linear multistep and Taylor expansion that are close to those we observe in Table 2.

Thus, we conclude that the transformer is likely using the matrix exponential method. This makes sense given the problem setting - both the linear multistep and Taylor expansion methods are only accurate for small Δt, while our bound of $\Delta t = U[0, 2\pi/\omega_0]$ violates this assumption for some timeseries. Still, it is remarkable that transformers use a known numerical method to model the undamped harmonic oscillator, and we can provide evidence for its use, although our experiments do not rule out the possibility of other methods being used in conjunction with the matrix exponential.

Criterion	Linear Multistep	Taylor Expansion	Matrix Exponential
1	0.66/0.51	0.67/0.25	**0.84**/**0.54**
2	0.73/**0.44**	0.74/0.39	**0.89**/**0.44**
3	0.42/0.15	0.53/0.11	**0.78**/**0.16**
4	0.44/X	0.44/X	**0.72**/X

Table 2. We summarize the evaluation of methods and criteria for the undamped/underdamped models. For each criteria, we list a single quantity for readability. Criterion 1 is the largest value in Fig. 5, criterion 2 is the correlation in Fig. 6, criterion 3 is the largest value in Fig. 7, and criterion 4 is the ratio in the legend of Fig. 8. The matrix exponential performs best across criteria.

3.4. Extension to the damped harmonic oscillator ($\gamma \neq 0$)

We want to understand the generality of our finding by extending our problem space to the damped harmonic oscillator, where $\gamma \neq 0$. We leave relevant details about our procedure to Appendix E, but in Table 2, we find our intermediate analysis performs much more poorly on the underdamped case than undamped. We describe possible explanations in Appendix E, but because of this we temper our finding from the undamped harmonic oscillator with caution about its generality.

4. Conclusions

After developing criteria for intermediates in the toy setting of linear regression, we find that transformers use known numerical methods for modeling the simple harmonic oscillator, specifically the matrix exponential method. We leave the door open for researchers to better understand the methods transformers use to model the damped harmonic oscillator and use the study of intermediates to understand how transformers model other systems in physics.

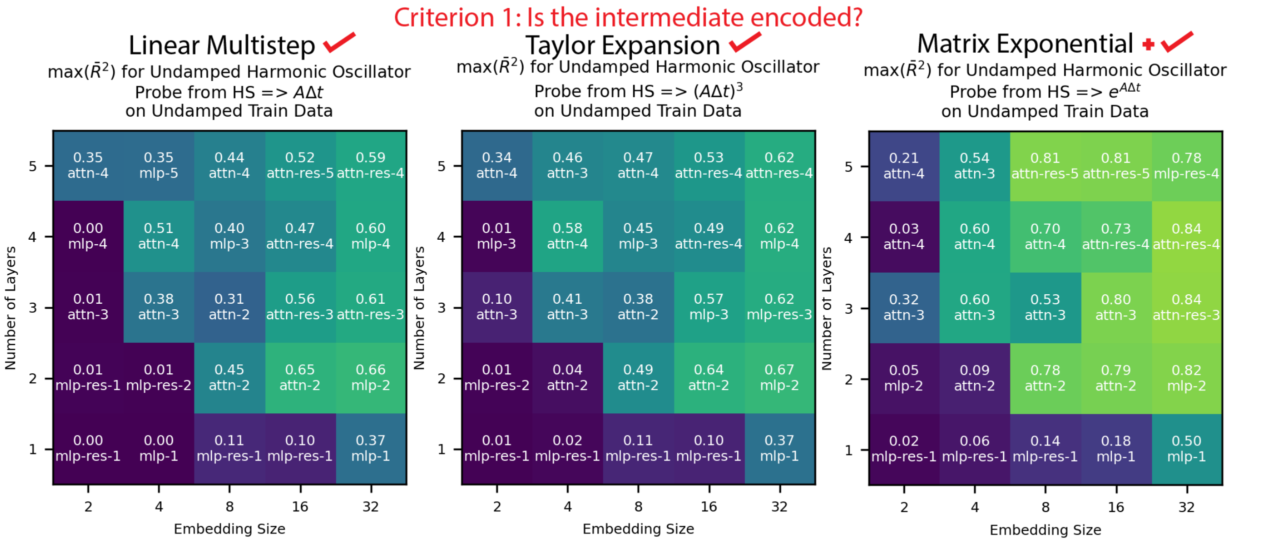

Figure 5. We analyze the intermediates of our undamped harmonic oscillator models, and find all three methods encoded, with the matrix exponential method best represented. This provides initial correlational evidence for all three methods.

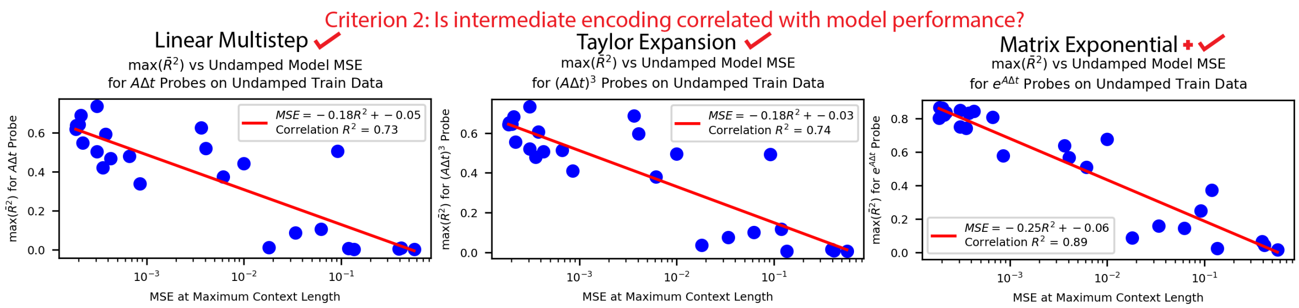

Figure 6. We find that better performing models have intermediates of all methods better encoded, but this correlation is strongest in magnitude and slope for the matrix exponential method. This is additional correlational evidence for all three methods.

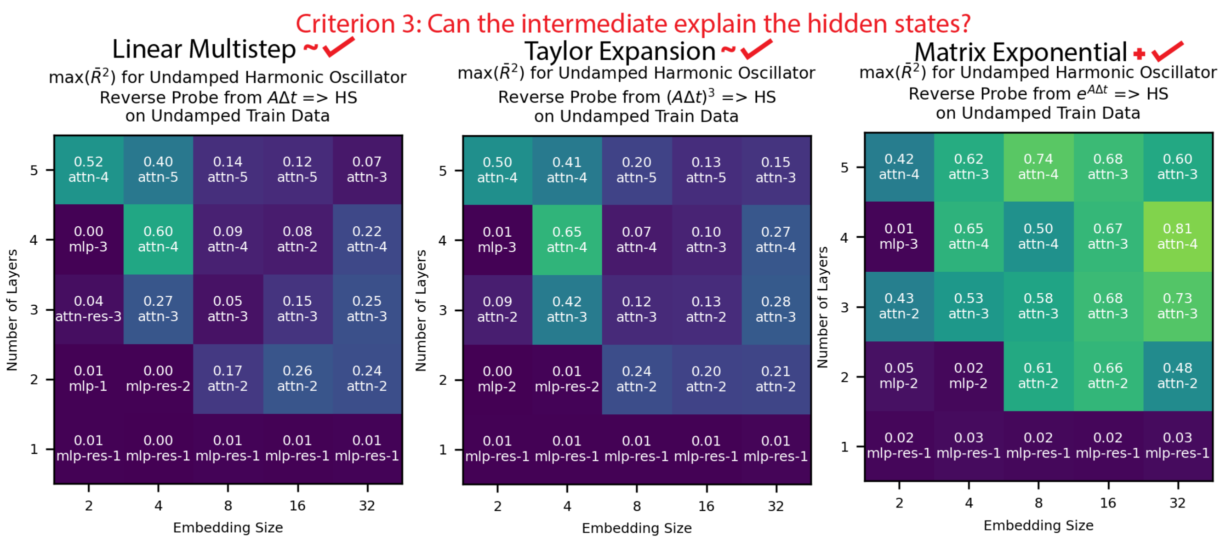

Figure 7. We find that the intermediates from all three methods can explain some variance in model hidden states, but the matrix exponential method is most consistent and successful by a wide margin.

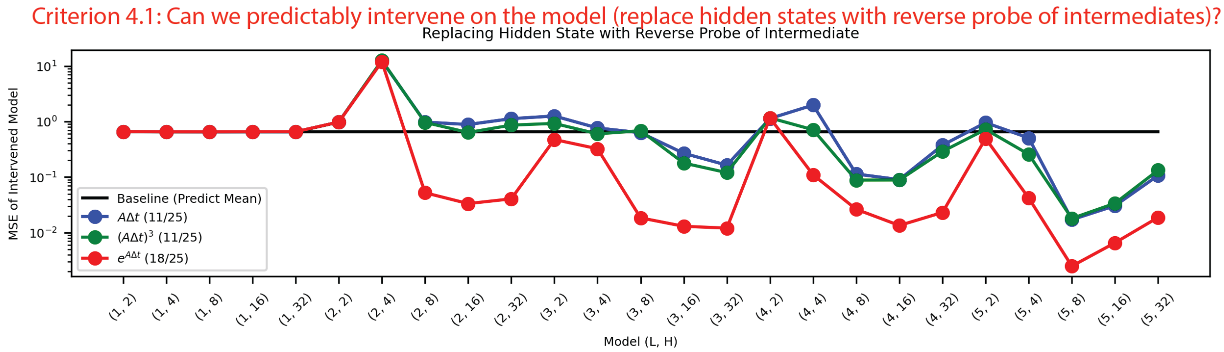

Figure 8. For each model and method, we replace the hidden state in Fig. 7 with the reverse probe of the intermediate. We see this intervention is consistently best performing for the matrix exponential method by an order of magnitude, and 18/25 models perform better than our baseline of guessing.

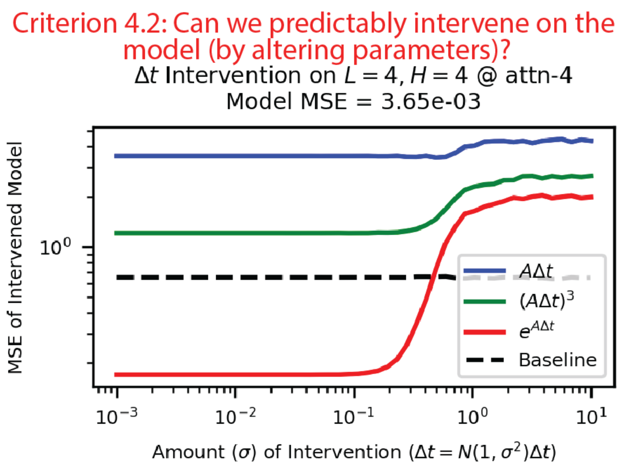

Figure 9. We vary the value of Δt used in the intermediates and use the reverse probes from Fig. 7 to generate hidden states from these intermediates. We perform this operation on the model with the best linear multistep/Taylor expansion ($L = 4, H = 4$) reverse probes, and find that the matrix exponential is consistently most robust to interventions. The baseline is if our model only predicted the mean of the dataset.

Limitations We analyze relatively small transformers with only one attention head and no LayerNorm. While we demonstrate strong results for the undamped harmonic oscillator, our results for the underdamped harmonic oscillator are more mild. We only use noiseless data.

References

[1] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2022.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[5] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Baker Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *ArXiv*, abs/2209.10652, 2022.

[6] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[7] Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663, 2022.

[8] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[9] Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *The Fortieth International Conference on Machine Learning*, 2023.

[10] Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *ArXiv*, abs/2305.01610, 2023.

[11] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*, 2023.

[13] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

[14] Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. *ArXiv*, abs/2306.17844, 2023.

[15] Toni J. B. Liu, Nicolas Boull'e, Raphael Sarfati, and Christopher J. Earls. Llms learn governing principles of dynamical systems, revealing an in-context neural scaling law. 2024.

[16] Joel A. Shapiro. *Classical Mechanics; Lagrange's and Hamilton's Equations*. Rutgers University, Piscataway, NJ, 1 edition, 2010.

[17] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *ArXiv*, abs/2211.15661, 2022.

[18] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *ArXiv*, abs/2208.01066, 2022.

[19] *Canonical Correlation Analysis*, pages 321–330. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[21] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv: Learning*, 2018.

[22] Ziming Liu and Max Tegmark. A neural scaling law from lottery ticket ensembling. *ArXiv*, abs/2310.02258, 2023.

[23] Abhilasha Ravichander, Yonatan Belinkov, and Eduard H. Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Conference of the European Chapter of the Association for Computational Linguistics*, 2020.

[24] *Runge–Kutta Methods*, chapter 3, pages 143–331. John Wiley & Sons, Ltd, 2016.

[25] *Linear Multistep Methods*, chapter 4, pages 333–387. John Wiley & Sons, Ltd, 2016.

[26] University of Victoria. Odes: Matrix exponentials. adapted for math 204 at the university of victoria. Accessed: 2024-05-16.

[27] Wes Gurnee and Max Tegmark. Language models represent space and time. *ArXiv*, abs/2310.02207, 2023.

[28] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *ArXiv*, abs/1610.01644, 2016.

[29] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification.

[30] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, 2023.

[31] Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16), April 2020.

[32] Ziming Liu and Max Tegmark. Machine learning conservation laws from trajectories. *Phys. Rev. Lett.*, 126:180604, May 2021.

[33] M. Cranmer, Sam Greydanus, Stephan Hoyer, Peter W. Battaglia, David N. Spergel, and Shirley Ho. Lagrangian neural networks. *ArXiv*, abs/2003.04630, 2020.

[34] Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Neural Information Processing Systems*, 2019.

[35] Subhash Kantamneni, Ziming Liu, and Max Tegmark. Optpde: Discovering novel integrable systems via ai-human collaboration. 2024.

[36] Steven L. Brunton. Notes on koopman operator theory. 2019.

Supplementary material

A. Related Work

Mechanistic interpretability Mechanistic interpretability (MI) as a field aims to understand the specific computational procedures machine learning models use to process inputs and produce outputs (5; 6; 13; 7; 8; 9; 10; 11; 12). Some MI work focuses on decoding the purpose of individual neurons (27) while other work focuses on ensembles of neurons (11; 12). Our work is aligned with the latter.

Algorithmic behaviors in networks A subset of MI attempts to discover the specific algorithms networks use to solve tasks by reverse engineering weights. For example, it has been demonstrated that transformers use the discrete Fourier transform to model modular addition (13), while additional work has found competing hypotheses for the specific algorithm transformers use on this task(14). Instead of reverse engineering weights, we make use of linear probing (28) to discover byproducts of algorithms represented internally by transformers. Studies have found that algorithms in models are potentially an "emergent" behavior that manifests with size (29; 30), which we also find.

AI & Physics Many works design specialized machine learning architectures for physics tasks (31; 32; 33; 34; 35), but less work has been done to see how well transformers perform on physical data out of the box. Recently, it was shown that LLMs can in-context learn physics data (15), which inspired the research question of this paper: how do transformers model physics?

B. Additional Results for Linear Regression

In Fig. 10, we find that our transformers are able to generalize to linear regression test examples with out-of-distribution data ($0.75 \leq |w| \leq 1$). In Fig. 11, 12 we see that smaller models do not have $\boldsymbol{w}$ encoded, while larger models often have $\boldsymbol{w}$ linearly encoded (with some quadratic encodings as well). In Fig. 13, we see that better performing models generally have better encodings, while worse performing models generally have worse encodings. We plot the relationship between ICL and encoding in Fig. 14.

C. Theory hub generalizes to other systems

We note that the theory hub we summarize in Table 1 is valid for all differential equations that can be written as $\dot{x} = \boldsymbol{A}\boldsymbol{x}$ if $\boldsymbol{A}$ is a constant matrix. This includes all homogenous linear differential equations with constant coefficients, and potentially non linear differential equations as well. Koopman operator theory allows nonlinear differential equations to be modeled as linear differential equations. Here is an example taken from (36):

Here, we consider an example system with a single fixed point, given by:

$$\dot{x}_1 = \mu x_1 \tag{32a}$$

$$\dot{x}_2 = \lambda(x_2 - x_1^2). \tag{32b}$$

For $\lambda < \mu < 0$, the system exhibits a slow attracting manifold given by $x_2 = x_1^2$. It is possible to augment the state x with the nonlinear measurement $g = x_1^2$, to define a three-dimensional Koopman invariant subspace. In these coordinates, the dynamics become linear:

$$\frac{d}{dt}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \mu & 0 & 0 \\ 0 & \lambda & -\lambda \\ 0 & 0 & 2\mu \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \quad \text{for} \quad \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \end{bmatrix}. \tag{33a}$$

For this nonlinear system, our theory hub in Table 1 is still relevant using

$$\boldsymbol{A} = \begin{bmatrix} \mu & 0 & 0 \\ 0 & \lambda & -\lambda \\ 0 & 0 & 2\mu \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \end{bmatrix}.$$

Thus, it is possible that the methods we've determined a transformer uses to model the simple harmonic oscillator extends to other, more complex systems.

D. Undamped Harmonic Oscillator Appendices

Data generation for the undamped harmonic oscillator We generate 5000 sequences of 65 timesteps for various values of $\omega_0, \Delta t, x_0$, and v_0. We range $\omega_0 = U[\frac{\pi}{4}, \frac{5\pi}{4}]$, $\Delta t = U[0, \frac{2\pi}{\omega_0}]$, $x_0, v_0 = U[-1, 1]$. The undamped harmonic oscillator is periodic so using a larger Δt is not useful. We also generate an out-of-distribution test set with $\omega_0 = U[0, \frac{\pi}{4}] + U[\frac{5\pi}{4}, \frac{3\pi}{2}]$ with the same size as the training set.

Additional results for undamped harmonic oscillator In Fig. 15, we see that models are able to learn the undamped harmonic oscillator in-context, even for values of ω_0 out of the distribution these models were trained on. We also plot the evolution of encodings for our various methods on the best performing undamped model in Fig. 16. We find that our choice of j for the Taylor expansion method has is mostly irrelevant for $j \leq 5$ in Fig. 17. With respect to criterion 4, in Fig. 18 we show that our intervention results are robust to which parameter we're intervening on ($\Delta t, \omega_0$, or both) for multiple models. We also generate synthetic hidden states from the matrix exponential intermediate and find that the values for criterion 1,3 for the other two methods

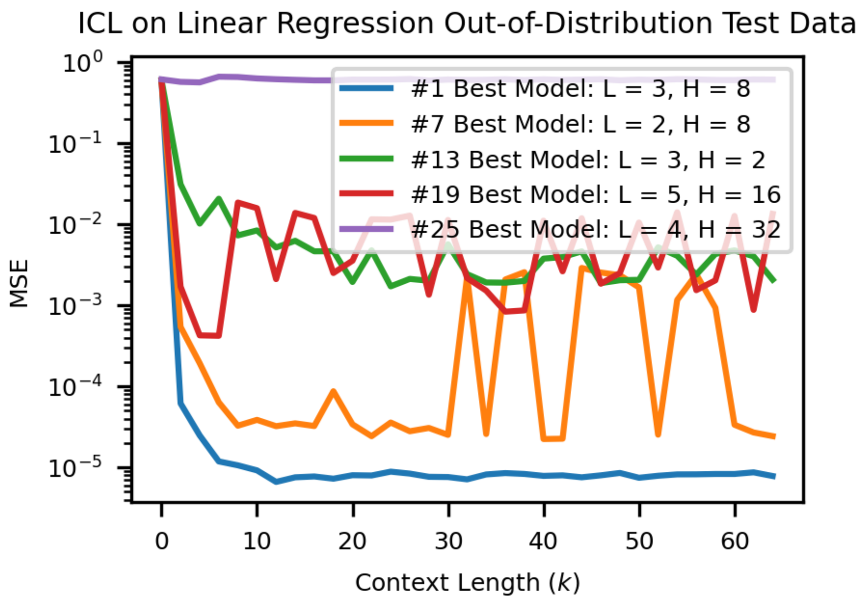

Figure 10. We find that linear regression models are able to generalize to out-of-distribution test data with $0.75 \leq |w| \leq 1$.

are potentially byproducts of the matrix exponential in Fig. 19, giving additional correlational evidence that the matrix exponential is the dominant method of the transformer.

E. Investigating the damped harmonic oscillator ($\gamma > 0$)

E.1. Mathematical setup

The damped harmonic oscillator has three well studied modes: underdamped, overdamped, and critically damped cases. The underdamped case occurs when $\gamma < \omega_0$, and represents a spring oscillating before coming to rest. The overdamped case occurs when $\gamma > \omega_0$, and represents a spring immediately returning to equilibrium without oscillating. The analytical equations for both cases are

Underdamped ($\gamma < \omega_0$)

$$\boldsymbol{x_k} = e^{-k\gamma\Delta t}\Big(x_0\cos(k\omega\Delta t) + \frac{v_0+\gamma x_0}{\omega}\sin(k\omega\Delta t)\Big)$$

$$\boldsymbol{v_k} = e^{-k\gamma\Delta t}\Big(v_0\cos(k\omega\Delta t) - \Big(\frac{v_0+\gamma x_0}{\omega}\gamma + \omega x_0\Big)\sin(k\omega\Delta t)\Big)$$

Overdamped ($\gamma > \omega_0$)

$$\boldsymbol{x_k} = \frac{e^{-k\gamma\Delta t}}{2}\Big((x_0 + \frac{v_0+\gamma x_0}{\omega})e^{k\omega\Delta t} + (x_0 - \frac{v_0+\gamma x_0}{\omega})e^{-k\omega\Delta t}\Big)$$

$$\boldsymbol{v_k} = \frac{e^{-k\gamma\Delta t}}{2}\Big((\omega-\gamma)(x_0 + \frac{v_0+\gamma x_0}{\omega})e^{k\omega\Delta t} - (\omega+\gamma)(x_0 - \frac{v_0+\gamma x_0}{\omega})e^{-k\omega\Delta t}\Big)$$

where $\omega = \sqrt{|\gamma^2 - \omega_0^2|}$. Note that the critically damped case ($\gamma = \omega_0$) is equivalent to $\lim_{\gamma\to\omega_0^-}$ of the underdamped case and $\lim_{\gamma\to\omega_0^+}$ of the overdamped case. Thus, we focus our study on the underdamped and overdamped cases, and visualize sample trajectories of both in Appendix Fig. 20.

E.2. Computational setup for the damped harmonic oscillator

We use an analogous training setup to the undamped harmonic oscillator. We generate 5000 sequences of 32 timesteps for various values of $\omega_0, \gamma, \Delta t, x_0$, and v_0 for both the underdamped and overdamped cases. For the underdamped and overdamped case, we range $\omega_0 = U[0.25\pi, 1.25\pi]$ and $\Delta t = U[0, \frac{2\pi}{13\omega_0}]$. We use this sequence length and bound on Δt to account for the periodic nature of the damped harmonic oscillator and also to ensure that the system does not decay to 0 too fast. For the underdamped case, we take $\gamma = U[0, \omega_0]$, and for the

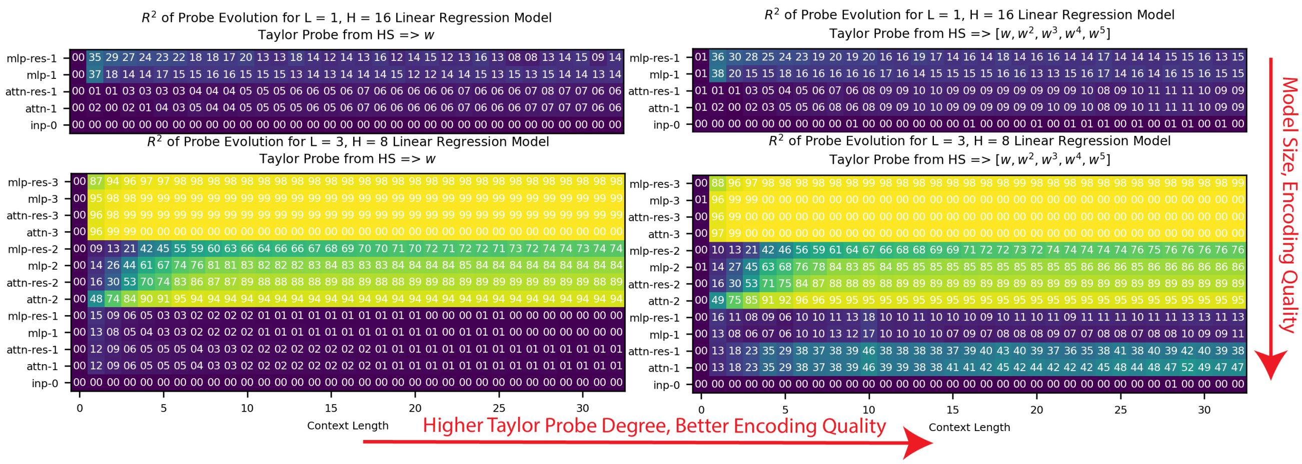

Figure 11. We plot the R^2 of Taylor probes for the intermediate $\boldsymbol{w}$ within models trained on the task $\boldsymbol{Y} = \boldsymbol{wX}$. We see that larger models have $\boldsymbol{w}$ encoded, often linearly, with little gain as we move to higher degree Taylor probes, while smaller models do not have $\boldsymbol{w}$ encoded.

overdamped case, $\gamma = U[\omega_0, 1.5\pi]$. We also generate an out-of-distribution test set following a similar process but using $\omega_0 = U[0, 0.25\pi] + U[1.25\pi, 1.5\pi]$.

In Fig. 20 we find that a transformer trained on *underdamped* data is able to generalize to *overdamped* data with only in-context examples! This is a surprising discovery, since a human physicist who is only exposed to underdamped data would model it with the analytical function in Section E.1. But this method would not generalize: the underdamped case uses exponential and trigonometric functions of $\gamma\Delta t$ and $\omega\Delta t$ respectively, while the overdamped case consists solely of exponential functions. We predict that our "AI Physicist" is able to generalize between underdamped and overdamped cases because it is using numerical methods that model the underlying dynamics shared by both scenarios.

E.3. Criteria are less aligned for the underdamped harmonic oscillator

We evaluate all methods for the underdamped harmonic oscillator on our criteria and summarize the evaluations in Table 2 and show relevant figures for criteria 1, 2, and 3 in Figures 21, 22, and 23 respectively. While we see moderate correlational and some causal evidence for our proposed methods, we note that there is a steep dropoff across criteria between the undamped and underdamped cases. We identify a few possible explanations for this discrepancy:

The transformer is using a method outside of the hypothesis space Because the intermediates explain so little of the hidden states even when combined (Fig. 23), we hypothesize that the transformer has discovered a novel numerical method or is using another known method outside of our proposed hypothesis space. This is more likely for the damped case because we decrease the range on Δt to avoid decay, which makes approximate numerical solutions more accurate. But why would it be doing this for the damped case and not the undamped case? For our damped experiments, we decrease the range on Δt so that the trajectory does not decay to 0 to quickly, but this also allows for approximate numerical methods to be more accurate, as demonstrated by the competitive performance of the linear multistep method with the matrix exponential method in Table 2. So it is possible our transformer is relying on another numerical method outside of our hypothesis space.

Natural decay requires less "understanding" by the transformer As the context length increases, damping forces the system to naturally decay to 0, so the transformer can use less precise methods to predict the next timestep. In Appendix Fig. 24, we see that the intermediates' encodings accordingly decay with context length, which possibly explains the underdamped case's diminished metrics.

More data for the transformer to encode With a nonzero damping factor γ, the intermediates we investigate in Table 1 have more non-constant values in their 2×2 matrices in the damped/undamped case: the linear multistep method has 3/2 values, the Taylor expansion method has 4/2 values, and the matrix exponential method has 4/3 unique values. The increased number of non-constant values could potentially make it more difficult to properly encode intermediates.

We leave the problem of understanding the damped harmonic oscillator to future work with intermediates.

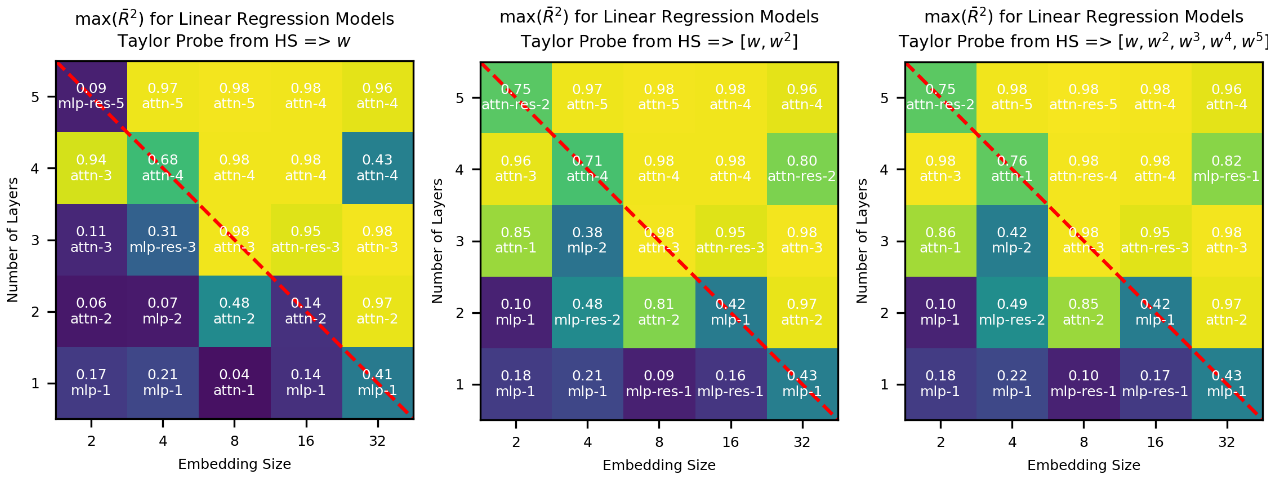

Figure 12. We calculate the mean of the R^2 of probes for $f(w)$ across all layers of the transformer and annotate each model with its highest mean score, $\max(\bar{R}^2)$. When $f(w)$ is linear (left) and quadratic (middle), we observe a striking phase transition of encoding based on model size, demarked by the red dashed line. If $\boldsymbol{w}$ is encoded, it is mostly encoded linearly, with the $(L, H) = (5, 2), (4, 32), (2, 8)$ models showing signs of a quadratic representation of $\boldsymbol{w}$. We do not see any meaningful gain in encoding when extending the Taylor probe to degree $n > 2$ (right). For models where $f(w)$ is well represented, it often happens in an attention layer. This is possibly because the attention layer aggregates all past estimates of $f(w)$ into an updated estimate.

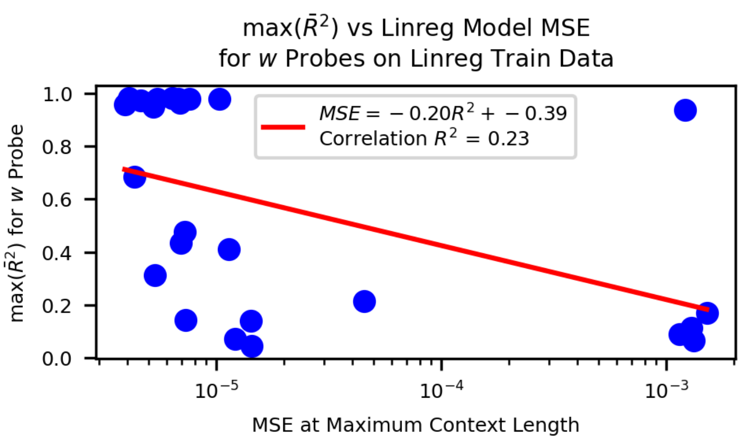

Figure 13. Better performing models generally have better encodings of $\boldsymbol{w}$, while worse performing models generally have worse encodings (other than one outlier in the top right)

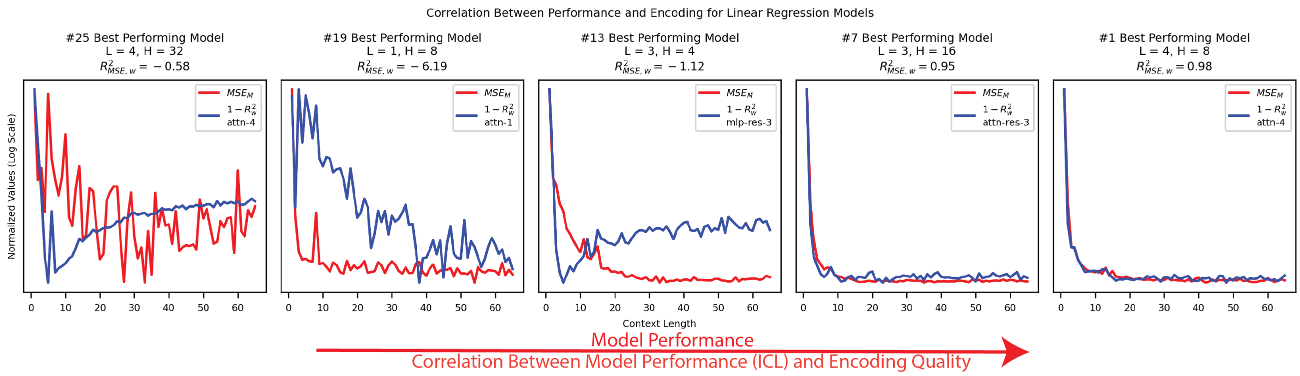

Figure 14. We test the correlation between model performance and the encoding of $\boldsymbol{w}$ on 5 of our 25 models of evenly spaced performance quality. We plot normalized values for the error of the encoding $(1 - R_w^2)$ in red and the mean squared error of the model (MSE_M) in blue. We find that the ability of the best performing models to in-context learn is highly correlated with their encoding of $\boldsymbol{w}$ $(R^2(MSE, w)$.

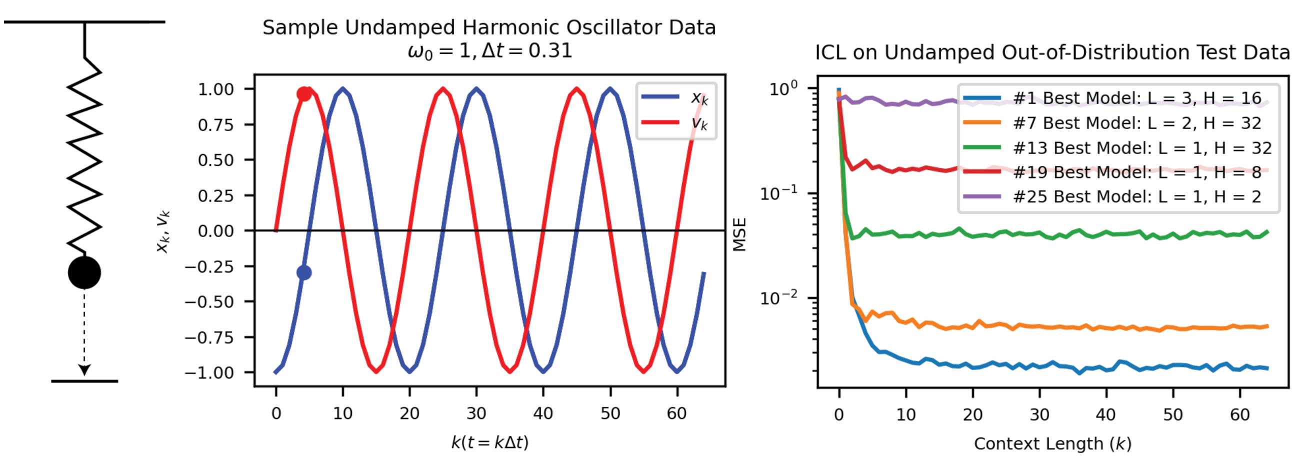

Figure 15. An intuitive picture for a simple harmonic oscillator is a mass oscillating on a spring (left). The trajectory of the SHO can be fully parameterized by the value of x, v at various timesteps (middle), and we find that models trained to predict undamped SHO trajectories are able to generalize to out-of-distribution test data with in-context examples (right).

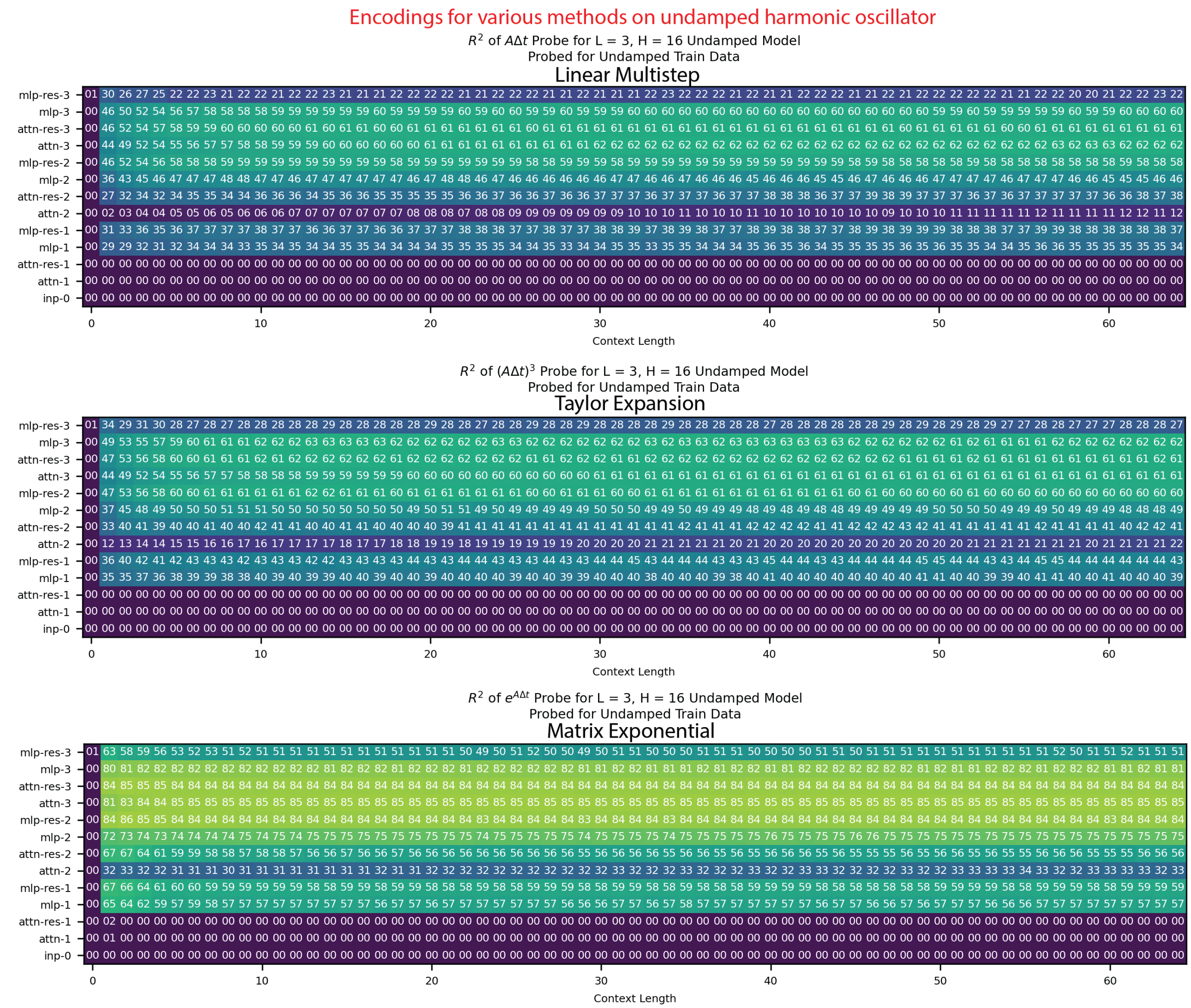

Figure 16. We visualize the evolution of encodings across all methods with context length for the best performing undamped model.

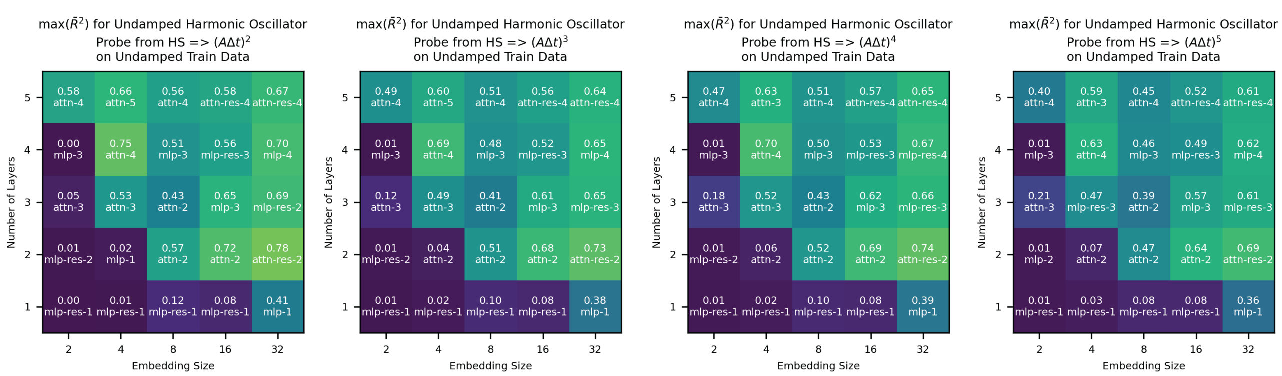

Figure 17. We find that our choice of j in the intermediate for the Taylor expansion method ($(A\Delta t)^j$ has little effect on our results or conclusions about the undamped harmonic oscillator (shown for criterion 1).

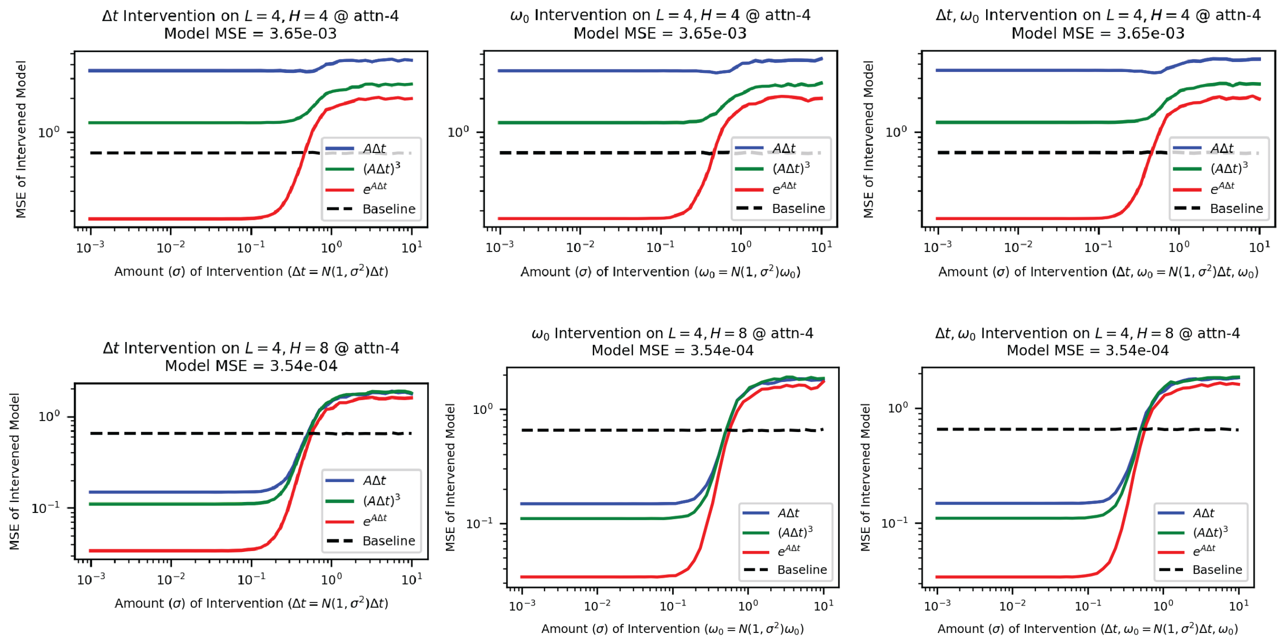

Figure 18. Regardless of which quantities we intervene on, our general results are robust for criterion 4 for the undamped harmonic oscillator. We perform interventions on the models with the best reverse probes for linear multistep/Taylor expansion ($L = 4, H = 4$) and matrix exponential ($L = 4, H = 8$).

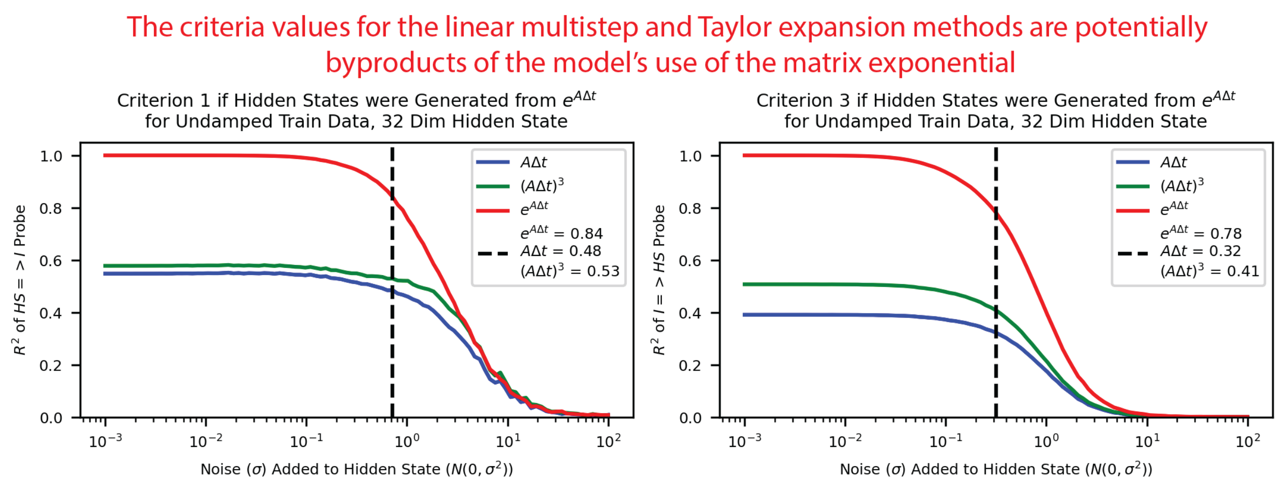

Figure 19. We generate synthetic hidden states from the matrix exponential intermediates and find that this naturally results in values for criterion 1,3 for the linear multistep and Taylor expansion methods that are close to those we observe in Table 2. This is correlational evidence that the matrix exponential method is potentially solely used by the transformer, and values for the other two methods are byproducts. These byproducts could arise because $e^{A\Delta t} = \sum_j (A\Delta t)^j / j!$

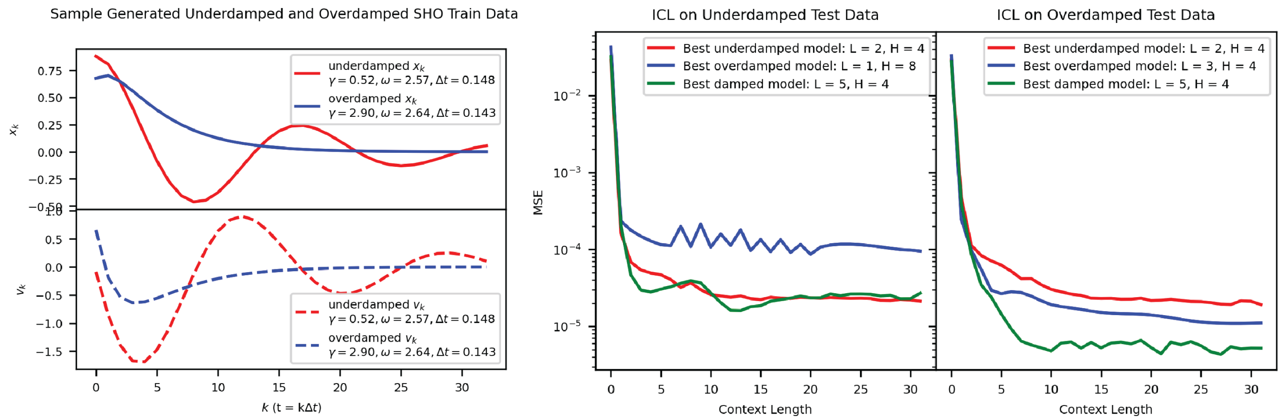

Figure 20. We generate data for underdamped and overdamped harmonic oscillators following the procedure detailed in Section 3, and visualize sample curves in the left most plot. From both the analytical equations and the plotted curves, we see that underdamped and overdamped data follow very different trajectories. Amazingly, on the right most plot we find that transformers trained on underdamped data generalize to overdamped data! This implies that our transformer is using a similar method to calculate both, otherwise this generalization would be impossible. We hypothesize that our "AI Physicist" is using one of the numerical methods from the undamped case. Note, that the "damped" oscillator was trained on equal parts underdamped and overdamped data.

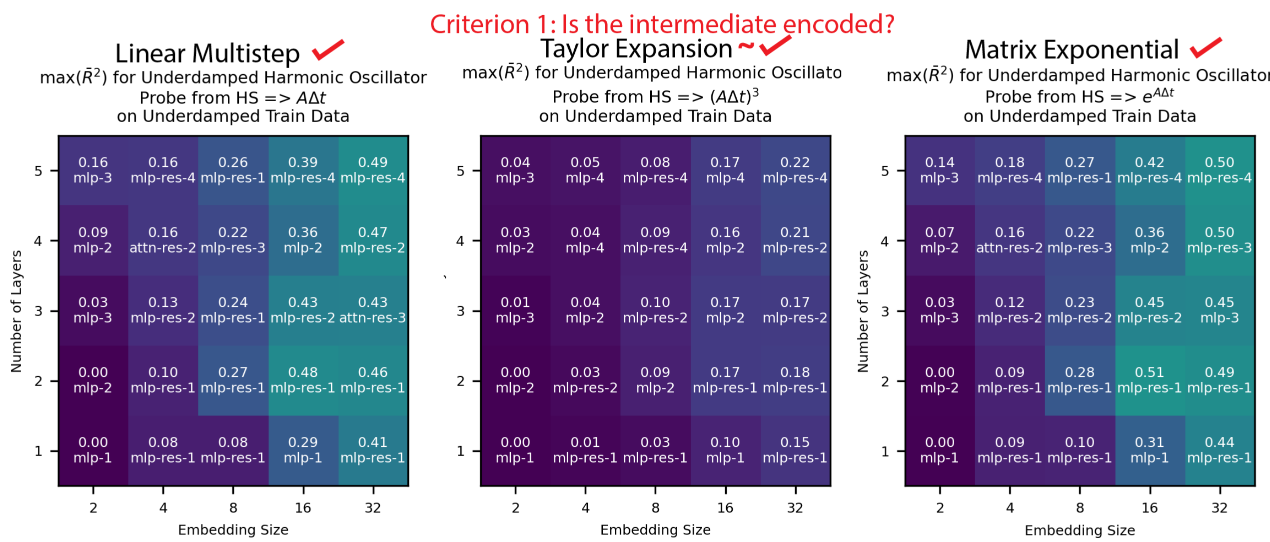

Figure 21. We observe that the intermediates for all three methods are encoded, but less than the undamped case in Fig. 5. The linear multistep is roughly as prominent as the matrix exponential method, which is also a departure from the undamped case.

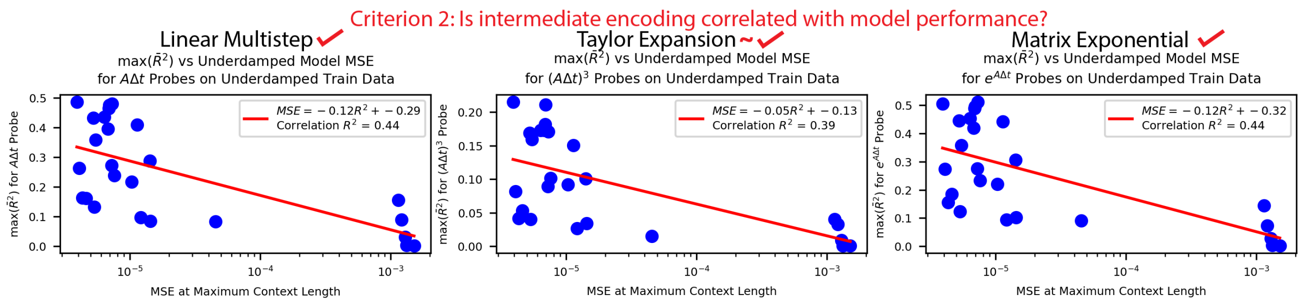

Figure 22. We see that generally, better performing models exhibit stronger encodings of intermediates, while worse performing models exhibit weaker encodings. These trends are not as strong as the undamped case, shown in Fig. 6. Like criterion 1 in Fig. 21, we see that the linear multistep method is competitive with the matrix exponential method.

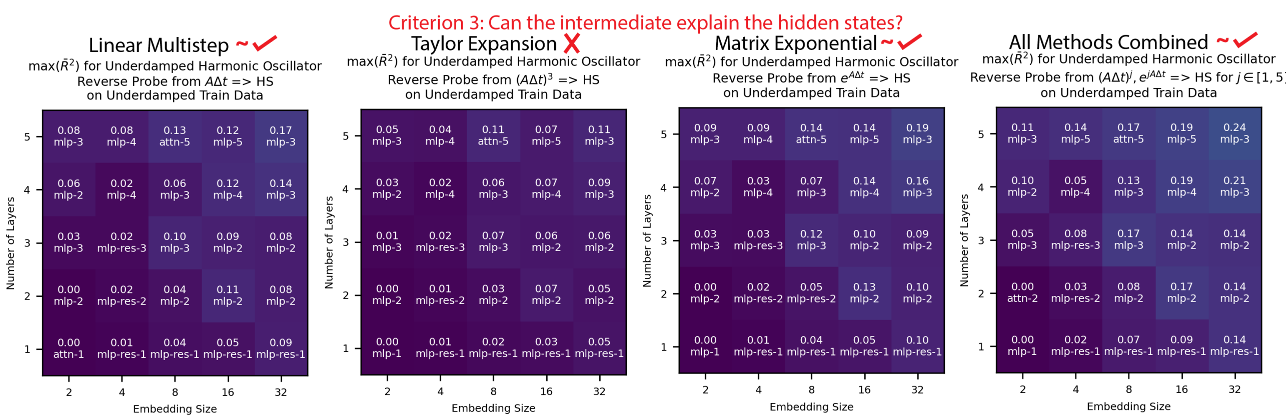

Figure 23. Multiple methods represent nontrivial amounts of variance in the hidden states, but even all methods combined (right) explain less than a quarter of the variance in the hidden states.

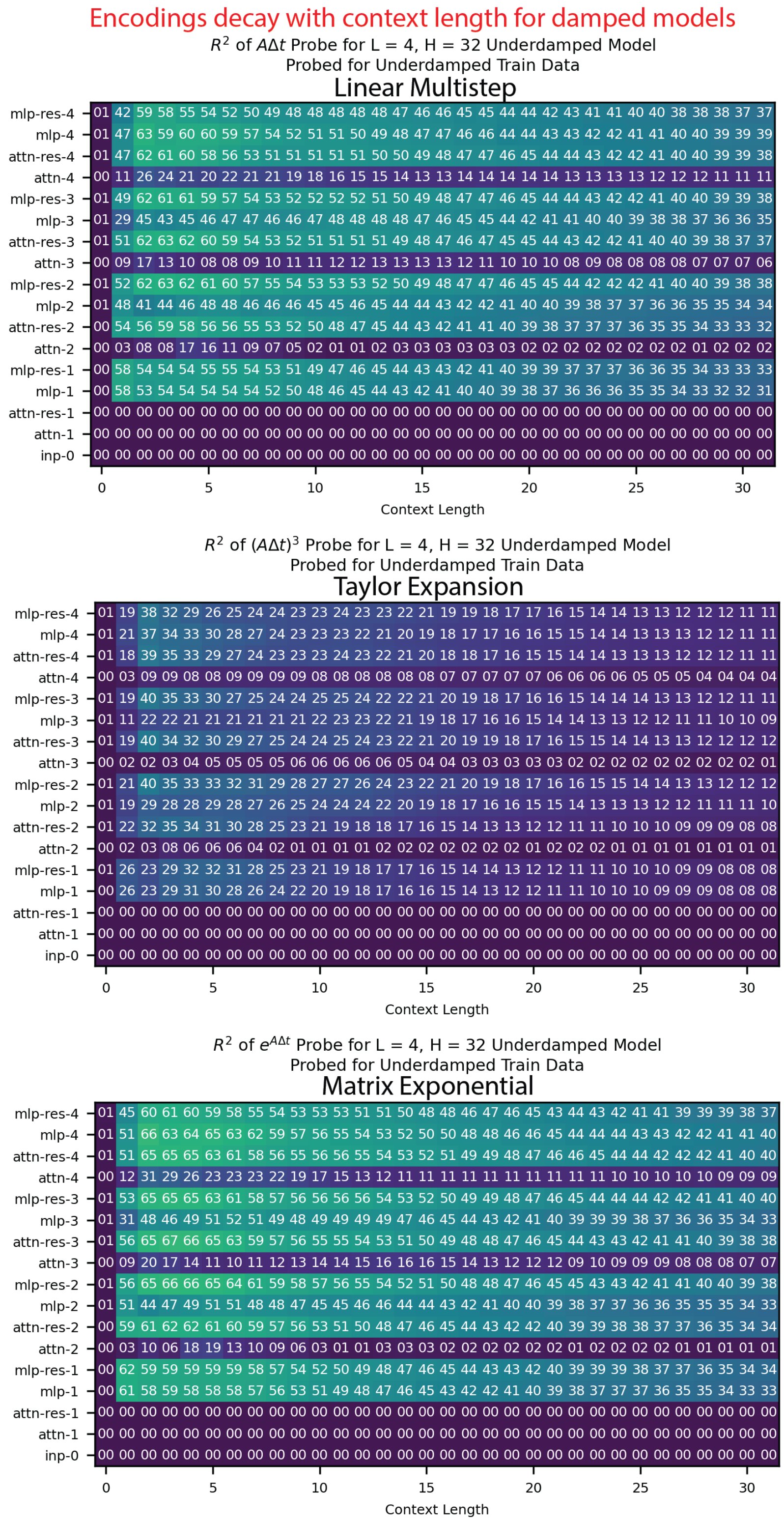

Figure 24. We see that the encoding strength of intermediates decays across all methods with context length. This similarly matches the natural decay to 0 of the damped harmonic oscillator, and is one potential explanation for why our methods are not as prominent in the damped vs undamped cases, for which the encoding quality does not decay with context length (Fig. 24). While this is a general observation across models, we visualize the $L = 4$, $H = 32$ model because it has the strongest encoding of intermediates from Fig. 21.

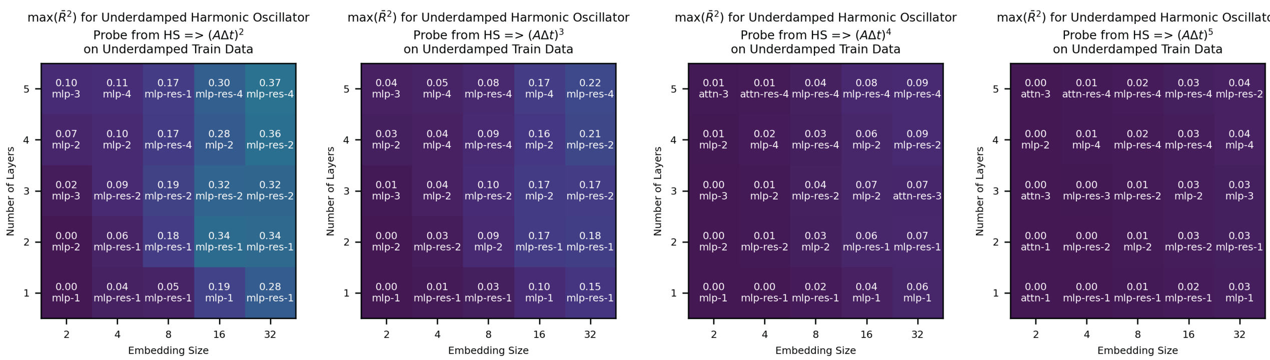

Figure 25. We find that our choice of j in the intermediate for the Taylor expansion method ($(A\Delta t)^j$ has a major effect on the encoding quality, unlike the undamped case visualized in Fig. 17. We see $j > 3$ is very poorly represented in the transformer, which implies that if the transformer was using the Taylor expansion for the underdamped spring, it would likely be of order $k = 3$ or less.