Measuring In-Context Computation Complexity via Hidden State Prediction, Additional Figures

1 Boring vs. Interesting Tasks for Specialized Models and Pre-Trained LLMs

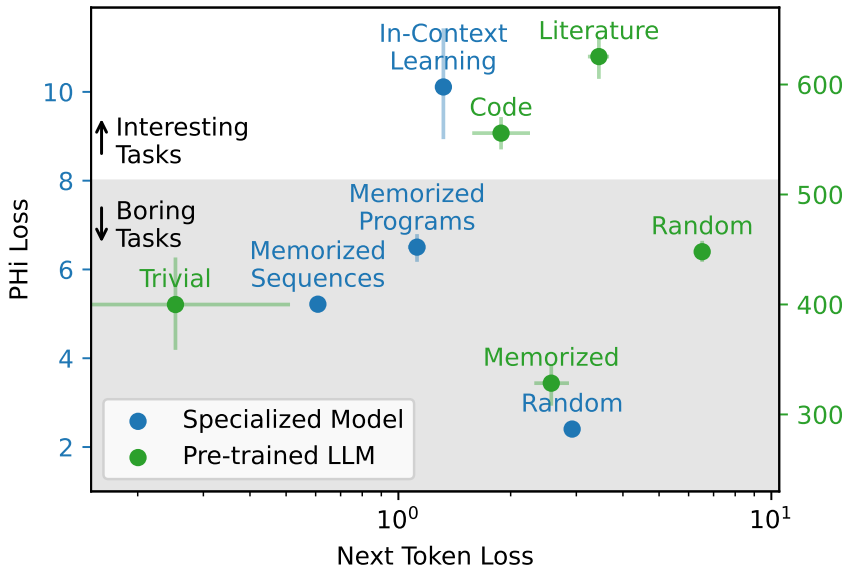

Figure 1: Interesting tasks, like in-context learning, or modeling of code and literature, exhibit high hidden state prediction (PHi) loss, while boring or trivial tasks, such as retrieving memorized sequences or modeling random structureless data, show low PHi loss. Next token loss provides no meaningful insight into task complexity. Results for a specialized transformer model (blue) and a pre-trained LLM (green), with PHi loss scales differing due to hidden state size. See Sections 3.1.1 and 3.2.1 for details.

2 Additional Results for Correct vs. Erroneous Rationales

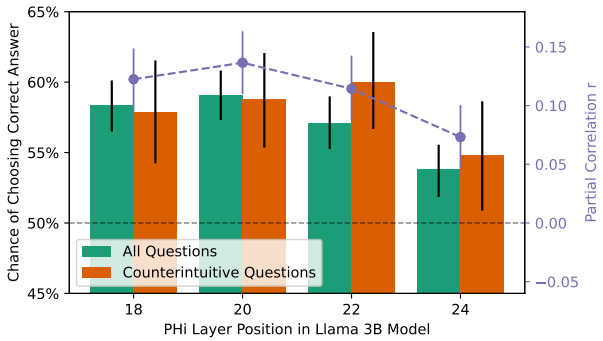

Figure 2: Rationales for the GSM-8k dataset. Same as Figure 8 in the submitted paper, but with additional confidence intervals for the partial correlation.

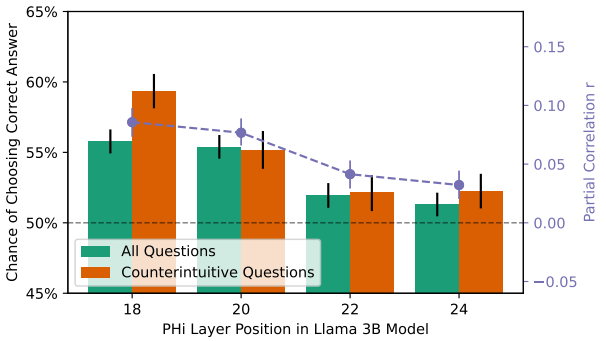

Figure 3: Similar to Figure 2, but for the MATH dataset. Chance of choosing the correct answer when selecting the one with the higher PHi loss between a correct and a wrong option. "Counterintuitive" are only those questions for which the answer with the lower next token prediction loss is wrong. In purple, the partial correlation r between PHi loss and answer correctness—controlled for next token prediction loss—is shown. Answers with high PHi loss are clearly more likely to be correct.

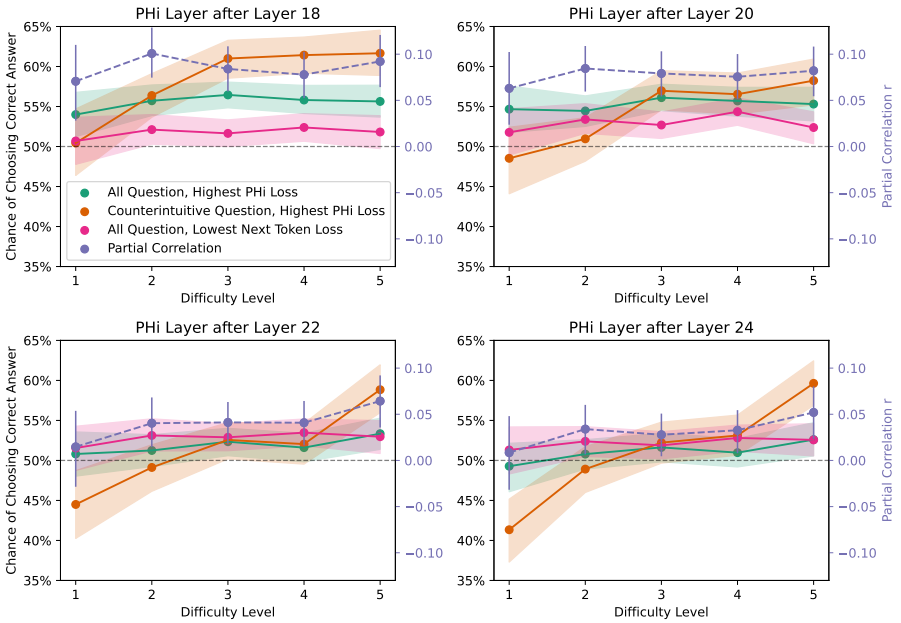

Figure 4: Detailed view of rationales generated for the MATH dataset, separated by the difficulty of the questions. Curves show (green) the chance of choosing the correct answer when selecting the one with highest PHi loss among all answer pairs, (orange) among the counterintuitive pairs, (pink) among all pairs when selecting the answer with lowest next token loss, and (purple) the partial correlation between PHi loss and answer correctness, controlled for next token loss. We see a strong correlation between the correctness of the rationale and PHi loss, especially for difficult counterintuitive questions. For easy counterintuintuitive question, this relationship does not exist.

3 PHi Layer Position in Fully Trained Transformer

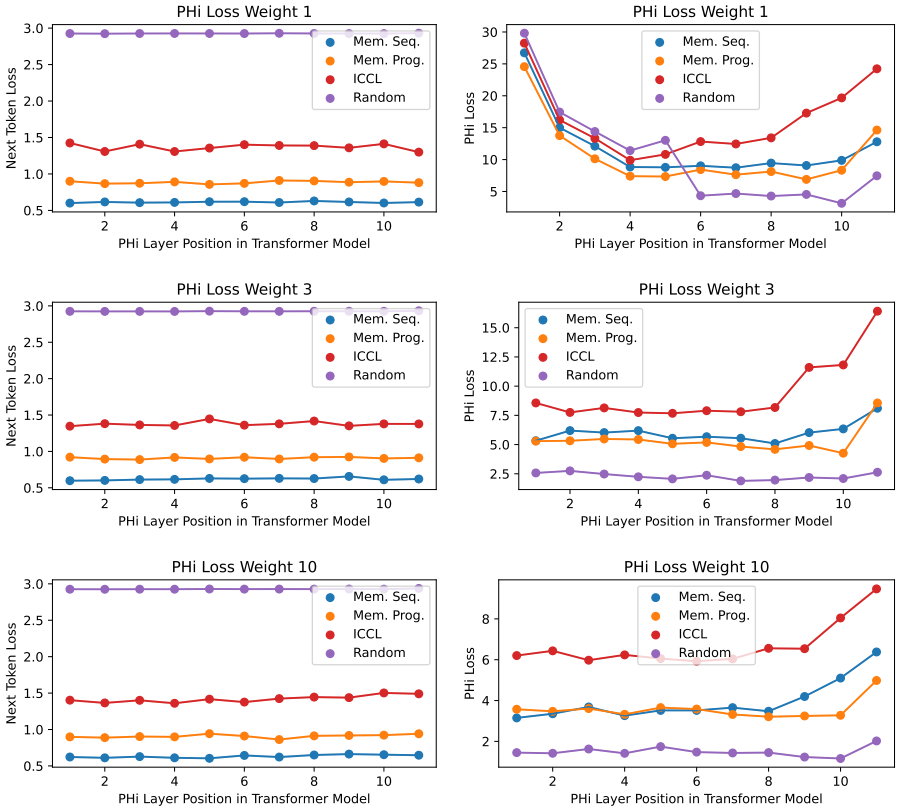

Figure 5: Experiments as described in Section 3.1.1, but with the PHi layer placed at different depths in the transformer model. Additionally, we vary the relative weight of the PHi loss compared to the next-token prediction loss. The loss at time step t (Equation 5) is given by $L(t) = L_{\text{NLL}}(t) + wL_{\text{PHi}}(t)$, where w is the PHi loss weight. The results in the main paper are reported for $w = 1$. However, we observe that higher weights also perform well, maintaining next-token prediction accuracy while improving robustness to the placement of the PHi layer.

4 Llama 3B without Hidden State Prediction

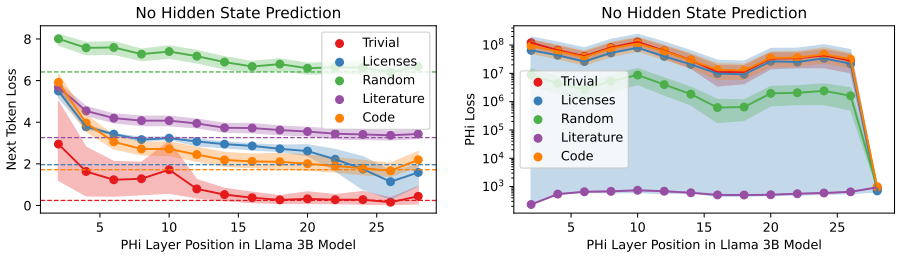

Figure 6: Experiments as described in Section 3.2.1, but using a PHi layer with a fixed unit Gaussian prior instead of a learned autoregressive prior. We observe consistently large PHi losses that do not correlate with task complexity. We tested various hyperparameter settings but were unable to obtain better results. This ablation highlights that hidden state prediction is essential for accurately estimating task complexity in pre-trained LLMs.

5 Llama 3B PHi Layer Training Curves

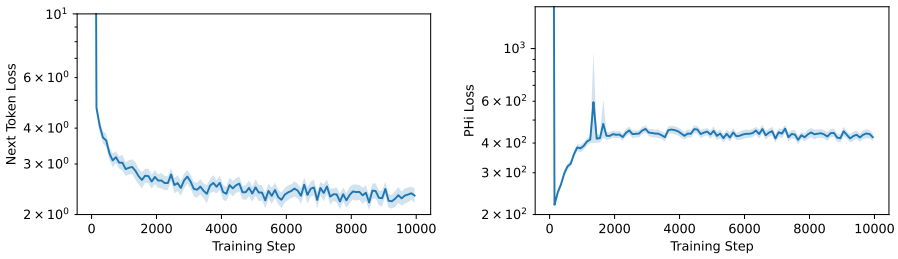

Figure 7: Typical training curves for a PHi layer placed after layer 20 in the pre-trained Llama 3B model. Training converges quickly, making it unnecessary to exceed 10,000 steps. Each training step processes approximately 4,000 natural language tokens. This run took 3.5 hours on a single NVIDIA RTX 3090 GPU.