SEQ-VCR:PREVENTING COLLAPSE IN INTERMEDIATE TRANSFORMER REPRESENTATIONS FOR ENHANCED REASONING

Md Rifat Arefin, Gopeshh Subbaraj, Nicolas Gontier, Yann LeCun, Irina Rish, Ravid Shwartz-Ziv, Christopher Pal

Md Rifat Arefin

PhD Candidate, Mila/University of Montreal

March 21, 2025

Outline

Background Work

- Representation Learning
- Kolmogorov Complexity
- Information bottleneck Principle

Motivation

- Representation Collapse
- Limitations on Multi-Step Reasoning

Our Solution: Seq-VCR (Sequential Variance Covariance Regularization)

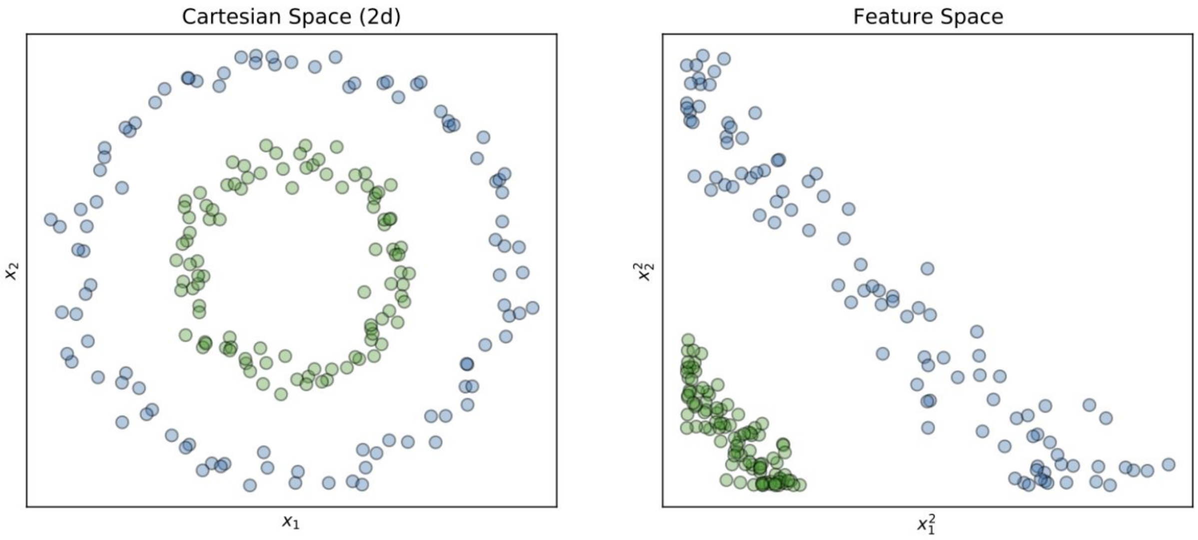
- Encourages Feature Diversity & Prevents Collapse
- Enhances Information Propagation Across Layers

Experimental Results

- Improving representational capacity
- Improving Performance Multi-Step Arithmetic Reasoning

Representation Learning

Representation learning is finding the good description of raw data into structured, meaningful abstractions that are easier to understand, process and reason about.



But what makes a **good** representation??

The Quest for Efficient Learning

Occam's Razor (14th century):

Among competing hypotheses, the one with the fewest assumptions (**simpler one**) should be preferred.

Aristotle's Posterior Analytics (4th Century BC):

The best demonstration is the one which is derived from the **fewer** postulates or hypotheses.

Kolmogorov Complexity – Measure of Simplicity

The complexity of data is the length of the shortest program(**compression**) that generates it.

- A datapoint like `123123123123` has **low complexity** (it can be described as “repeat 123 four times”).
- A random sequence like `9s4jX2#@!k5` has **high complexity** (no compressible pattern).

Information Bottleneck Principle *(Tishby et. al.)*

When **X**: input, **Z**: latent representation **Y**: output,

Learning Objective:

$$L = -\underbrace{I(Z; Y)}_{\text{Relevance}} + \beta \underbrace{I(X; Z)}_{\text{Compression}}$$

- I(X;Z): The mutual information between the **input** X (the previous tokens in LLMs) and the **latent representation** Z, which measures how much **relevant information** from the input is retained in Z.
- I(Z;Y): The mutual information between the Z and the **output** Y (the next token), which measures how much information in Z is relevant for predicting the output.
- β: A tradeoff parameter that controls the balance between **compression** (minimizing I(X; Z)) and **relevance** (maximizing I(Z;Y)).

Measuring Compression/Representation Collapse

Entropy $H(X)$ and $H(Z)$, serves as an upper bound to MI:

- $I(X; Z) = H(X) - H(X \mid Z) \leq H(X)$
- $I(X; Z) = H(Z) - H(Z \mid X) \leq H(Z)$

- **Compression occurs when H(Z) decreases across layers**.

Matrix/Prompt Entropy as Representation Collapse (Giraldo et al., Skean et al.)

- **Matrix-Based Entropy**, is a tractable surrogate for Rényi's α-order entropy, computed using eigenvalues of a similarity kernel.
- We use **Linear Kernel,** aligning with the **linear representation hypothesis** *(Park et al., 2024)*, that LLMs encode **high-level concepts** (truth, honesty etc.) in **linearly separable directions**.
- As a linear kernel K, we can use either the **Gram matrix** ($Z^{(l)}Z^{(l)T}$) or the **Covariance matrix** ($Z^{(l)T} Z^{(l)}$), where $Z^{(l)}$ represents token-level representations from the **l-th** layer with dimension **d**. Both matrices share the same nonzero eigenvalues, ensuring that the entropy calculation remains consistent regardless of the choice of kernel.

Eigenvalue of K

$$S_\alpha = \frac{1}{1-\alpha} \log \left[\sum_{i=1}^{T} (p_i)^\alpha \right] \qquad p_i = \frac{\lambda_i(K)}{\sum_i \lambda_i(K)}$$

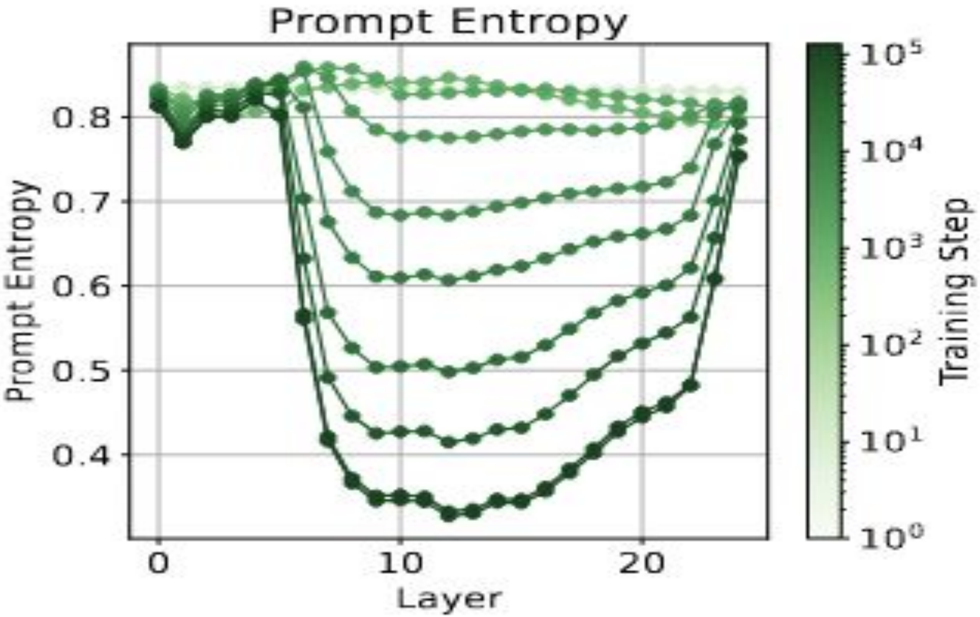
α -> 1, this reduces to Shannon entropy.

- The entropy captures how well information is spread **along linear directions** in the representation space.
- If representations are well-distributed, entropy is high; if they collapse into a few dominant directions, entropy is low.

Training Dynamics and Prompt Entropy *(Skean et. al)*

Pre-training dynamics of Pythia 410M parameter model:

- **Representation Collapse in Intermediate Layers**: As pre-training progresses, intermediate layers exhibit increased representation collapse.
- **Information Bottlenecks**: Collapse restricts the flow of information across layers, potentially limiting the model's capacity to integrate knowledge effectively.
- **Task-Specific Implications**:
    a. While beneficial for certain tasks requiring compact representations,
    b. It may hinder multi-step reasoning tasks that require deeper information propagation.



Limitations of LLM: Multi-Step Reasoning

Who was the president of the United States when the Apollo 11 moon landing took place?

Part 1: When the Apollo 11 moon landing took place?

Answer: **1969**

Part 2: Who was the president of United States in 1969?

Answer: **Richard Nixon**

Final Answer: **Richard Nixon**

Tokenwise Complexity Imbalance on Multiplication Task

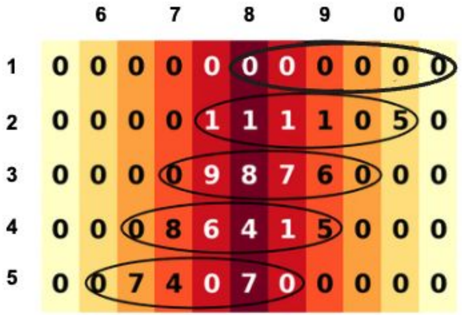
Challenges in nxn Multiplication:

Multi-Step Computation:

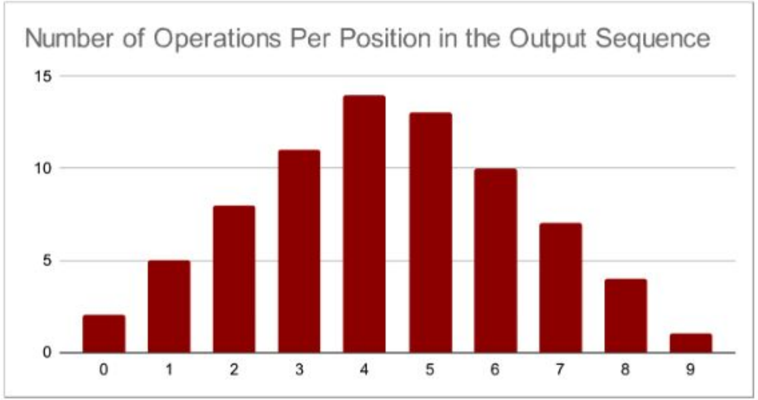
- The task requires **storing intermediate results**, demanding a **deeper model** for accurate processing.

Complexity Imbalance:

- Middle token requires more interactions with tokens and better representations



Position	0	1	2	3	4	5	6	7	8	9
Mult	1	2	3	4	5	5	4	3	2	1
Sum	0	1	2	3	4	4	3	2	1	0
Carry	1	2	3	4	5	4	3	2	1	0
Total	**2**	**5**	**8**	**11**	**14**	**13**	**10**	**7**	**4**	**1**

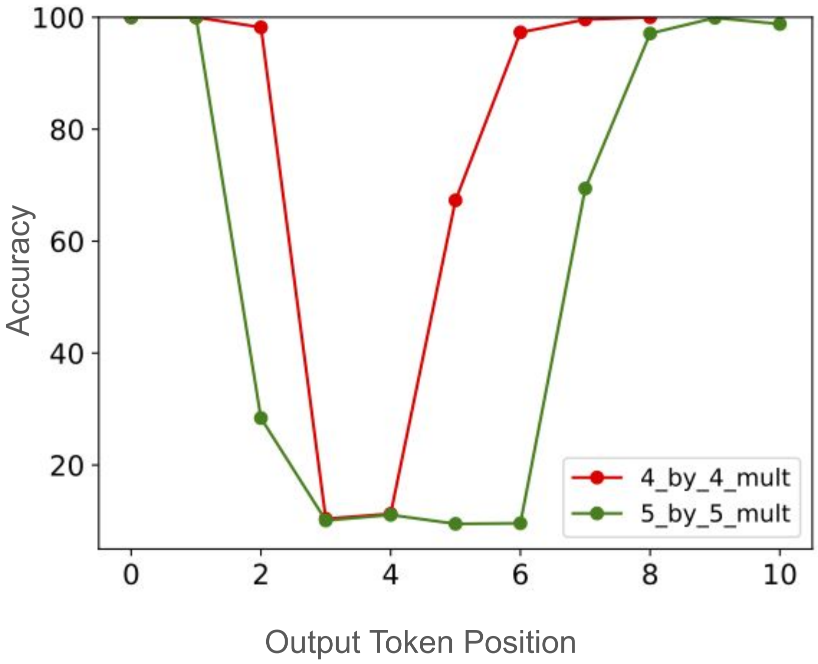


U-Shape like Token Accuracy on Multiplication

Finetune GPT2-small on nxn integer Multiplication without Chain of Thought:

- We observer U-shape like token-wise accuracy distribution

- Model can predict the peripheral tokens but fails on the middle ones.



Common solutions for multi-step reasoning

- **Increasing Model representation capacity**: increasing model size

  GPT2 < GPT3.5 < GPT4

- **Inference time compute**: decomposition with CoT prompting

Arithmetic Expression

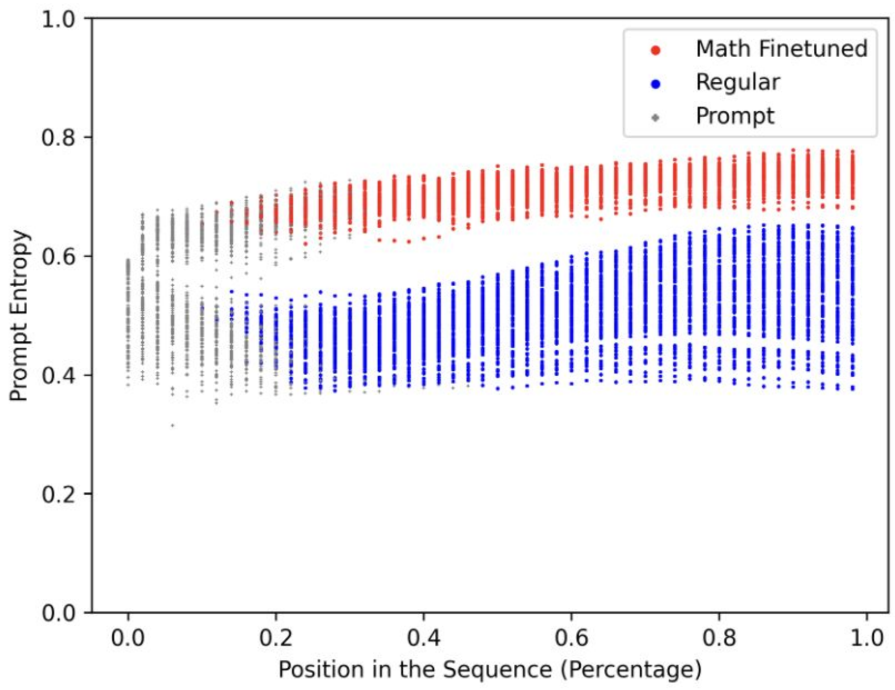
Input:
$(7+5) \div (6+4\times 3-2\times 7) =$

Output:
$$\begin{aligned} & 12 \div (6+4\times 3-2\times 7) \\ = & \ 12 \div (6+12-2\times 7) \\ = & \ 12 \div (18-2\times 7) \\ = & \ 12 \div (18-14) \\ = & \ 12 \div 4 \\ = & \ 3 \end{aligned}$$

Reasoning Traces and Prompt Entropy *(Skean et. al)*

Chain of Thought Reasoning Traces of Qwen 2.5 and Qwen 2.5-Math Models on GSM-8K:

- The base model (**Qwen 2.5**) exhibits greater **prompt compression**.
- The fine-tuned model (**Qwen 2.5-Math**) maintains **higher entropy**, indicating greater **information retention**.



Things required for Multi-step Reasoning

1) Inference time compute

   → CoT tokens or pause tokens

   We propose to use pause tokens as a proxy to add more inference time compute for the model

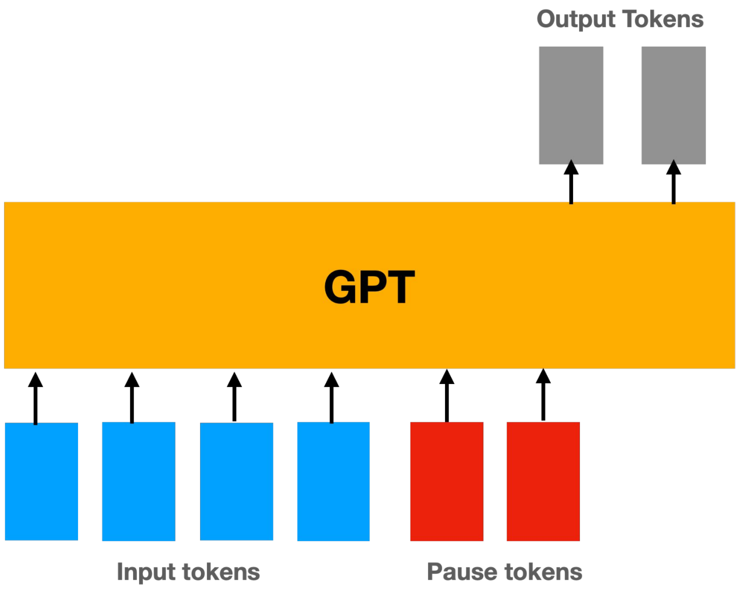
2) More Representation Capacity

   → Increased model size or Entropy regularization: Sec-VCR

   We aim to increase the representation capacity of **same size** models by reducing their representation collapse.

More Compute Through Pause tokens (Goyal et al.)

<question> </pause_start> <pause> <pause> </pause_end> <answer>

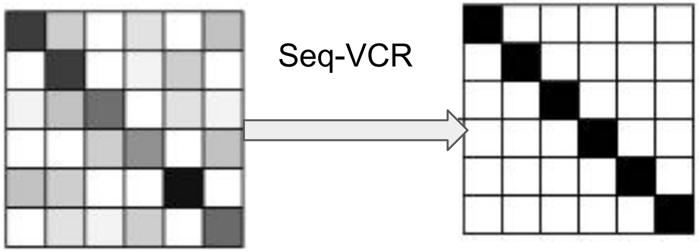


- Pause tokens are like randomly initialized tokens repeated and appended with input tokens

Increase Representation capacity: Seq-VCR

- **Extending VICReg(Bardes et al.) for LLM Representations**:
  - **VICReg** (Variance-Invariance-Covariance Regularization) was originally proposed for vision models.
  - We extend **VICReg** for LLMs to improve representation learning by diagonalizing the Covariance Matrix**.**
- **Covariance Diagonalization and Entropy**:
  - Prior work (Shwartz-Ziv et al.) shows that making the covariance matrix **diagonal** increases the entropy of representations.
  - Encouraging decorrelated features prevents representation collapse, promoting more efficient information propagation

$$L_{\text{Seq-VCR}} = \frac{1}{T \times d} \sum_{i=1}^{T} \sum_{k=1}^{d} \left(\lambda_1 \underbrace{\max(0, 1 - \sqrt{\mathbf{C}_{i,k,k} + \eta})}_{\text{Variance Term}} + \lambda_2 \underbrace{\sum_{k \neq \hat{k}} (\mathbf{C}_{i,k,\hat{k}})^2}_{\text{Covariance Term}} \right)$$



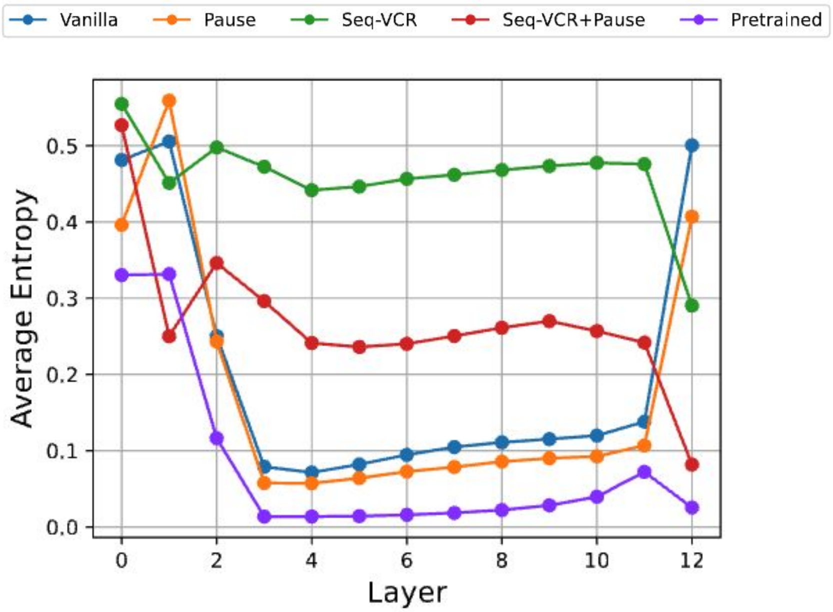
Where:

- **C** is the covariance matrix across the batch dimension of shape dxd
- λ_1 and λ_2 are regularization coefficients.
- η is a small constant for numerical stability.
- **Variance Term** ensures feature variance does not collapse.
- **Covariance Term** encourages decorrelation between features.

Configurations

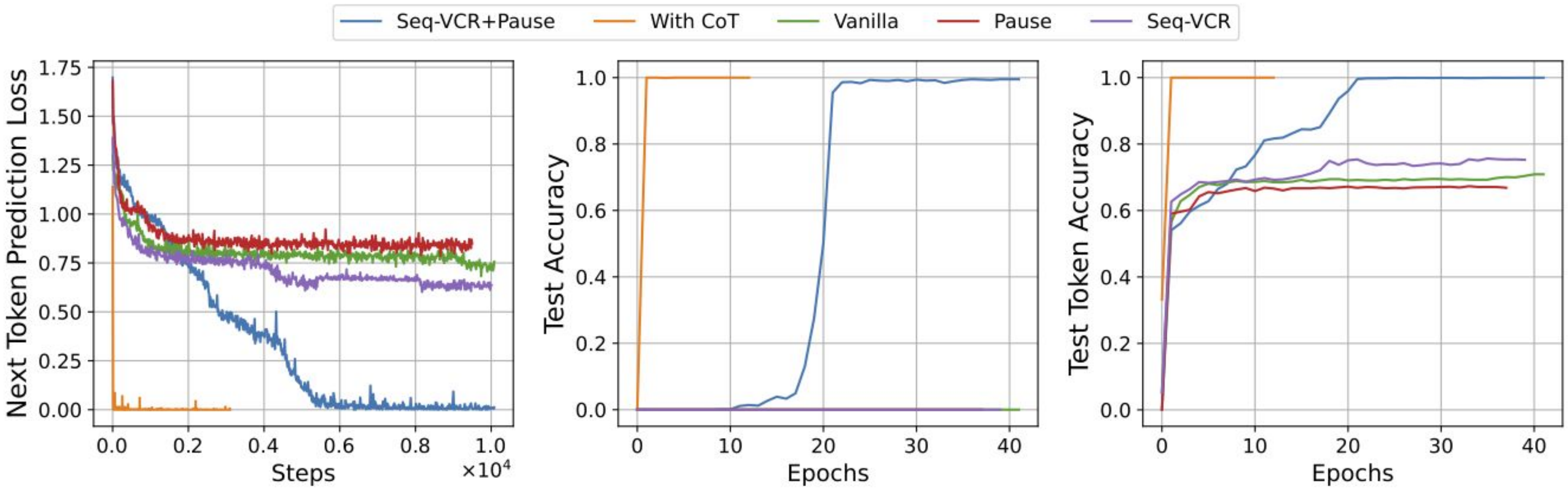
- **Vanilla:** Standard training/finetuning without regularization or pause/CoT tokens.
- **Pause:** Inserting pause tokens in the input sequence, no regularization.
- **Seq-VCR:** Applying Seq-VCR regularisation, no pause tokens.
- **Seq-VCR + Pause:** Combining Seq-VCR with pause tokens.
- **Pretrained:** Pre-trained language Model.
- **CoT:** Training/Finetuning with with CoT tokens.

Improving Representation Collapse



(b) Fine-tuning GPT-2 Small on 5×5 *digit Multiplication*

Finetuning Dynamics on Multiplication



Training Loss

Exact match Accuracy

Token-wise Accuracy

Results on Multiplication

Model	Configuration	4x4 Mult	5x5 Mult
GPT-3.5	With CoT	0.43	0.05
	No CoT	0.02	0.00
GPT-4	With CoT	0.77	0.44
	No CoT	0.04	0.00
GPT-2 Small	With CoT	1.0	1.0
	Vanilla	0.25	0.0
	Pause	0.28	0.0
	Seq-VCR	0.52	0.0
	Seq-VCR + Pause	0.992	0.995

- **Accuracy (exact match) on 4 × 4 and 5 × 5 digits Multiplication Tasks. GPT-3.5 and GPT-4 results are taken from [Deng et al.](#)) which are produced by 5-shot prompt**

Training from Scratch on more dataset

```
Arithmetic Expression

INPUT    7 + (12 ÷ 4) × 3² − 5 + 8
       = 7 + 3 × 3² − 5 + 8
       = 7 + 3 × 9 − 5 + 8
CoT    = 7 + 27 − 5 + 8
       = 34 − 5 + 8
       = 29 + 8
OUTPUT   37
```

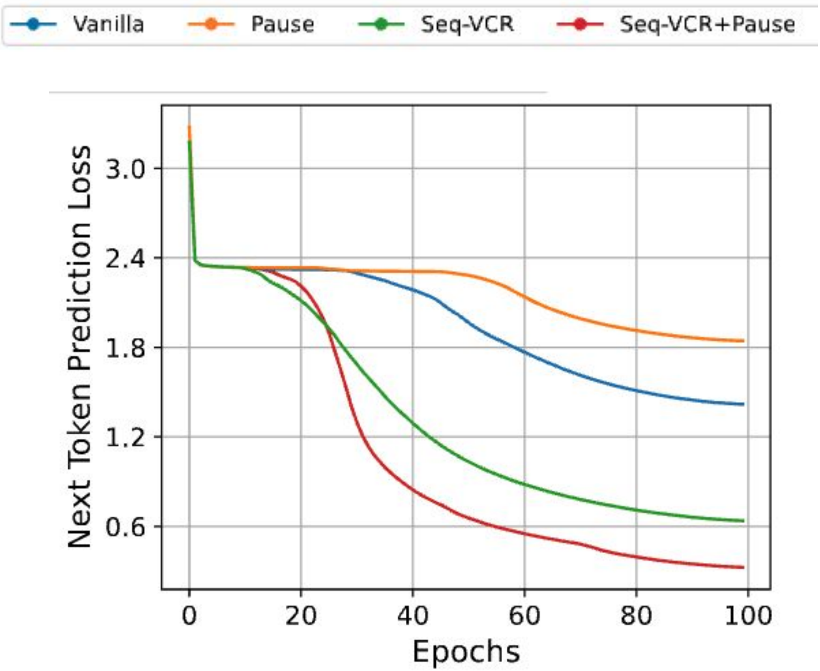
```
Longest Integer Subsequence

INPUT      [3, 10, 2, 1, 20]
        Initial dp[] Array: [1, 1, 1, 1, 1]
        Idx 0: dp = [1, 1, 1, 1, 1]
        Idx 1: dp = [1, 2, 1, 1, 1]
CoT     Idx 2: dp = [1, 2, 1, 1, 1]
        Idx 3: dp = [1, 2, 1, 1, 1]
        Idx 4: dp = [1, 2, 1, 1, 3]
OUTPUT     max(dp) = 3
```

Training Dynamics on Arithmetic Expression Task

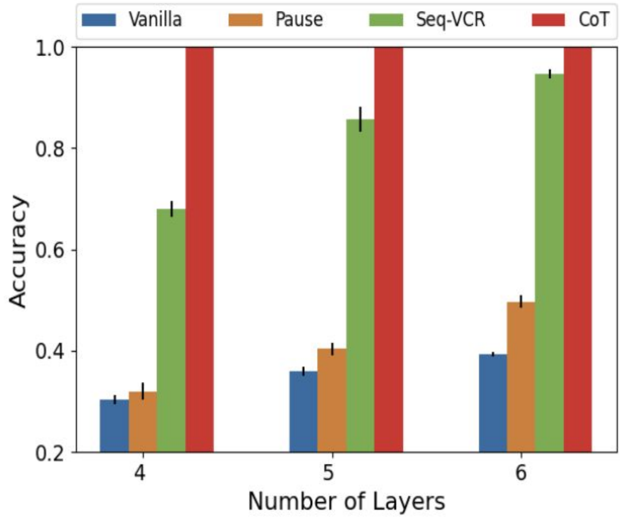
Dynamics of model training from scratch on Arithmetic expression task

- We observe sharp transition with Seq-VCR

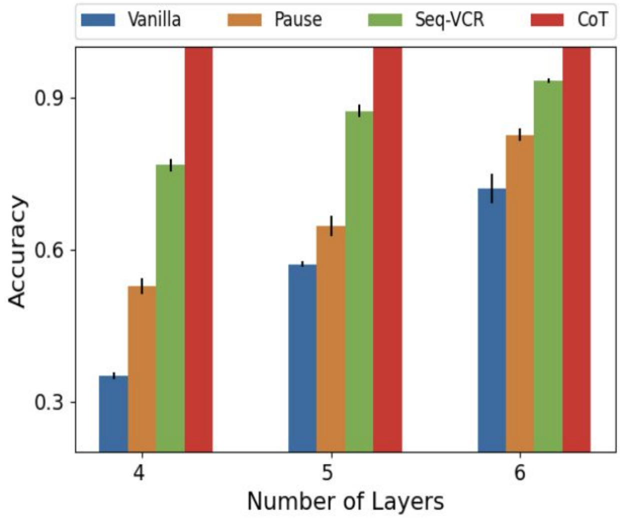


Varying # Layers

- We see consistent gains across number of layers



(a) Test accuracy on 6 operator Arithmetic Expression

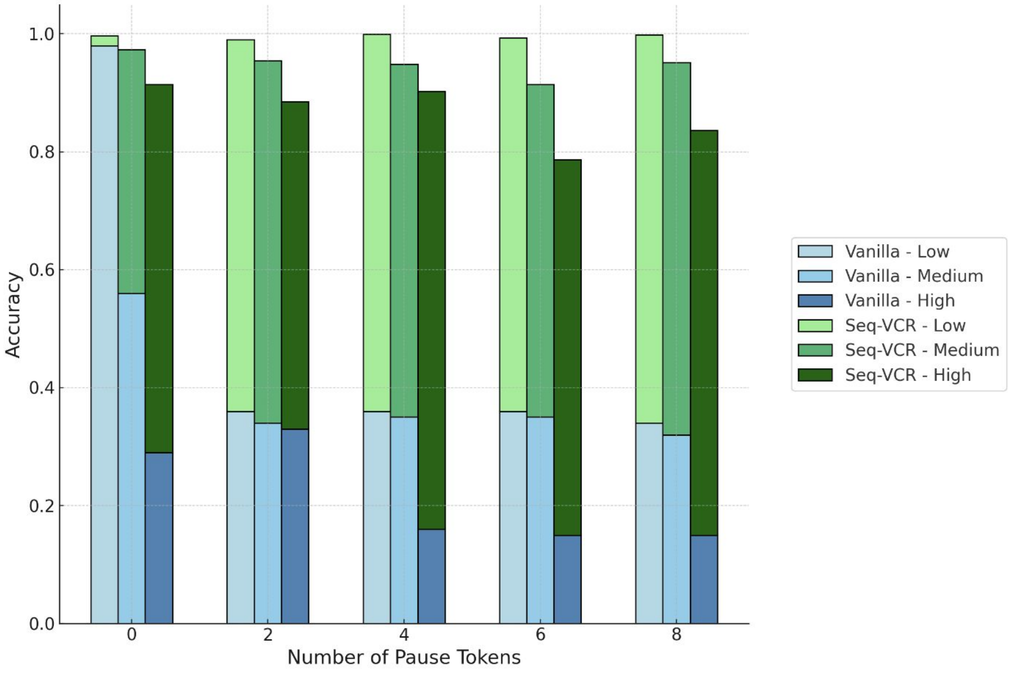


(b) Test accuracy on LIS Dataset with 100 Input Sequence Length

Varying Task Complexity

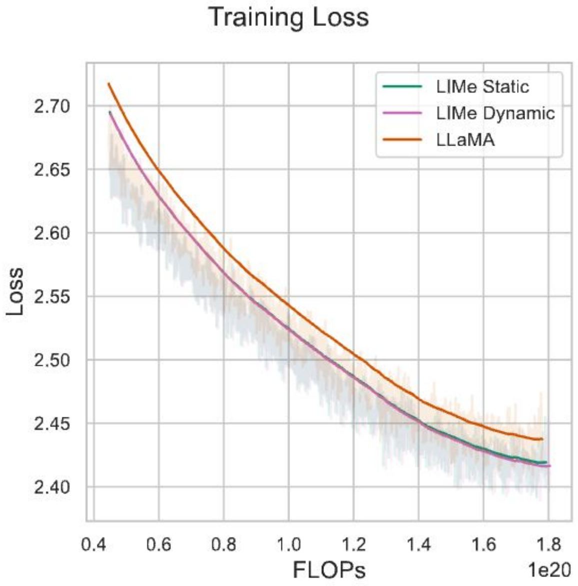
Varying Pause Tokens and Comparing Vanilla vs Seq-VCR + Pause Tokens for different Task Complexities.

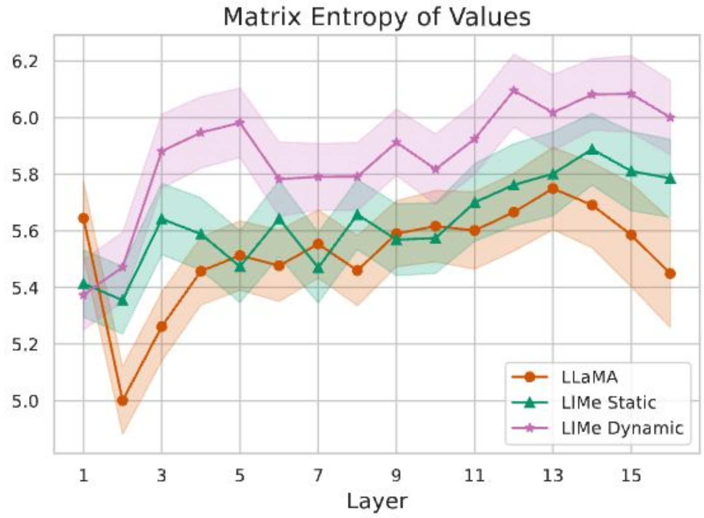
- Low, High, Medium refer to 4, 5, 6 arithmetic Operators respectively
- We using pause tokens with regular training is not useful.



Increase Representation Capacity in Pre-training *(Gerasimov et. al.)*

- Store the past layers activations to attend over





Conclusion & Future Directions

Key Findings

- Matrix-based entropy provides a robust framework for analyzing LLM representations.
- Representation collapse during pre-training restricts information flow, impacting multi-step reasoning.
- **Seq-VCR** regularization enhances representation quality and mitigates collapse.

Next Steps

- Investigate **Seq-VCR** interventions to improve LLMs' general reasoning ability.
- Explore **pretraining improvements** to develop models with higher representation capacity.

Thank You